Supplementary Materials for

Health and population effects of rare gene knockouts in adult humans with related parents

Vagheesh M. Narasimhan[1], Karen A. Hunt[2†], Dan Mason[3†], Christopher L. Baker[4†], Konrad J. Karczewski[5,6†], Michael R. Barnes[7], Anthony H. Barnett[8], Chris Bates[9], Srikanth Bellary[10], Nicholas A. Bockett[2], Kristina Giorda[11], Christopher J. Griffiths[2], Harry Hemingway[12,13], Zhilong Jia[7], M. Ann Kelly[14], Hajrah A. Khawaja[7], Monkol Lek[5,6], Shane McCarthy[1], Rosie McEachan[3], Anne O'Donnell-Luria[5,6], Kenneth Paigen[4], Constantinos A. Parisinos[2], Eamonn Sheridan[3], Laura Southgate[2], Louise Tee[14], Mark Thomas[1], Yali Xue[1], Michael Schnall-Levin[11], Petko M. Petkov[4], Chris Tyler-Smith[1], Eamonn R. Maher[15,16], Richard C. Trembath[2], Daniel G. MacArthur[5,6], John Wright[3], Richard Durbin[1‡*], David A. van Heel[2‡*]

correspondence to: rd@sanger.ac.uk or d.vanheel@qmul.ac.uk

This PDF file includes:

Materials and Methods
SupplementaryText
Figs. S1 to S8
Tables S1 to S8

Data Files S1 to S3 as Supporting Online Material.

Materials and Methods

Exome sequencing of 3222 Pakistani-heritage adult individuals living in the UK

Subjects and phenotypes

Birmingham adult Pakistani-heritage subjects comprised 130 subjects from a Birth Cohort study in Birmingham, UK(27), 471 adult healthy, and 892 adult type 2 diabetes subjects from Birmingham and Coventry as part of the UK Asian Diabetes Study(28) - a total of 1493 DNA subject samples, resulting in (after all quality control steps) 1060 exome sequenced subject samples. Approval was from the Birmingham East, North and Solihull Research Ethics Committee and from the South Birmingham Research Ethics Committee.

Born in Bradford (BiB) is a longitudinal multi-ethnic birth cohort study aiming to examine the impact of environmental, psychological and genetic factors on maternal and child health and wellbeing(29, 30). The full BiB cohort recruited 12,453 women during 13,776 pregnancies between 2007 and 2010. Ethical approval was granted by Bradford Research Ethics Committee. BiB adults studied here comprised 2490 DNA subject samples from pregnant women (>95% self-stated Pakistani-heritage) ascertained at an antenatal clinic visit in Bradford, UK(29), resulting in (after all quality control steps) 2162 exome sequenced subject samples. The first 1570 samples were of unselected parental relatedness status and included 554 with self-stated first cousin parental relatedness, the remaining samples were all of self-stated first cousin parental relatedness. We deliberately included (before the quality control stage) duplicate samples (taken at different antenatal visits for each of multiple pregnancies). We also recalled selected subjects for a second sample for genotype validation by a different method.

Exome sequencing and sequence-level quality control

Genomic DNA was extracted from peripheral blood, and quality confirmed by agarose gel electrophoresis (requiring high molecular weight) and picogreen assay. Samples were genotyped for identity checking by either Sequenom (San Diego, USA) assay (26 common autosomal variants, and 4 gender markers) or Fluidigm (San Francisco, USA) assay (22 common autosomal variants, and 4 gender markers). Samples where DNA-based gender and stated gender were mis-matched were excluded.

Genomic DNA (approximately 1 ug) was fragmented to an average size of 150 bp and subjected to DNA library creation using established Illumina paired-end protocols. Adapter-ligated libraries were amplified and indexed via PCR. A portion of each library was used to create an equimolar pool comprising 8 indexed libraries. Each pool was hybridised to SureSelect RNA baits (Human All Exon V5, Agilent Technologies, Santa Clara, USA) and sequence targets were captured and amplified in accordance with manufacturer's recommendations. Enriched libraries were used for 75 base paired-end sequencing (HiSeq 2000, Illumina, San Diego, USA) following manufacturer's instructions. Sequencing was performed to an expected ~40x read-depth, since the primary aim was to identify homozygous genotypes in autozygous regions.

Sample filtering

Stated duplicates were obtained from the BiB cohort, which took repeat samples when the same individuals registered at the Bradford hospital maternity ward for multiple pregnancies. A total of 149 individuals who registered for 2 pregnancies and 4 individuals who had registered for 3 pregnancies were recorded. From these duplicate pairings a total of 153 pairs had sequence data. Duplicate checks were performed using bcftools (version: 1.1-113-gd991f3f used throughout) gtcheck -G1 to compute the pairwise discordance between samples. 1000 random bi-allelic SNP sites were chosen out of a set of markers that were of at least MAF>0.05 in both the BiB dataset as well as the 1000 Genomes Project Phase3 release set, and had at least a mean sequencing depth of 20. Duplicates could be clearly separated from other samples based on number of discordant genotypes between each pair of sample. A duplicate sample was then removed from each pair so that there were unique samples for subsequent analyses.

To confirm sample identity and exclude laboratory mix-ups we compared the genotype calls made using Sequenom or Fluidigm as described above with those from our sequencing data. We removed 30 samples for which there was a genotype discordance of more than 30%.

Principal components analysis (PCA) to determine individual ethnicity was performed by merging bi-allelic SNPs from the current dataset with the 1000 Genomes Project phase 3 release set. Bi-allelic SNP sites that were of at least MAF>0.05 in each of the datasets were taken to define a set of markers used to perform the PCA analysis. PCA was first performed with 1000 Genomes Project data using all populations, and samples from the current dataset projected onto that reference. Pakistani-heritage subjects could clearly be defined, distinct from other South Asian ancestries on component 2 and 3 of the PCA and a region of the plot containing the majority of the samples was defined, leading to the removal of 294 samples that plotted outside the region (**fig. S2**). These 294 contained 20 samples with a 10-fold enrichment in the number of singleton sites which support ancestry other than the main cohort of British Pakistani subjects. Finally, using VerifyBamID(31) (freemix parameter > 0.03), we removed 24 samples that were predicted to have contamination between the samples.

Alignment and BAM processing

Following generation of raw reads, these were converted to BAM format using illumina2BAM and lanes de-multiplexed so that the tags were isolated from the body of the read, decoded, and used to separate out each lane into lanelets containing individual samples from the multiplex library and the PhiX control. Reads corresponding to the PhiX control were mapped and used with WTSI's spatial filter program to identify reads from other lanelets that contained spatially oriented INDEL artefacts and mark them as QC fail. Reads were aligned with BWA-MEM to the GRCh37+decoy reference genome used by the 1000 Genomes project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz). PCR and optically duplicated reads were marked using Picard MarkDuplicates, and after manual QC passing data was deposited with the EBI-EGA.

In order to ensure the quality of the large quantity of BAMs produced for the project, an automated quality control system was employed to reduce the number of data files that required manual intervention. This system was derived from the one originally

designed for the UK10K project(14) and used a series of empirically derived thresholds to assess summary metrics calculated from the input BAMs. These thresholds included: percentage of reads mapped; percentage of duplicate reads marked; various statistics measuring indel distribution against read cycle and an insert size overlap percentage. Any lane that fell below the "fail" threshold for any of the metrics were excluded; any lane that fell below the "warn" threshold on a metric would be manually examined; and any lane that did not fall below either of these thresholds for any of the metrics was given a status of "pass" and allowed to proceed into the later stages of the pipeline.

Passed lanelets were then merged into BAMs corresponding to sample's libraries and duplicates were marked again with Picard after which they were then merged into BAMs for each sample. Next, for each sample we re-aligned reads around known and discovered indels followed by base quality score recalibration using GATK (version: v3.3-0 used throughout). Lastly samtools 'calmd' was applied and indexes were created. Known indels for realignment were taken from Mills-Devine(32) and 1000 Genomes Project Phase 1(33) low coverage set, available from the 1000 Genomes ftp site. Known variants for base quality score recalibration were taken from dbSNP 137.

Variant calling

Two variant call-sets, one with the Genome Analysis Toolkit(34, 35) (GATK) HaplotypeCaller and one with samtools/bcftools(36), were made from the 4,353 samples that passed QC measures to this point. Calling was restricted to the Agilent V5 exome bait regions +/- a 100bp window on either end. The following parameters were used:

GATK HaplotypeCaller [GATK version: v3.3-0]:

Single-sample genome VCF (gVCF) files were created using GATK HaplotypeCaller run in gVCF mode on each of the sample BAM files using parameters:

-T HaplotypeCaller -R hs37d5.fa -variant_index_type LINEAR -variant_index_parameter 128000

--disable_auto_index_creation21478889_and_locking_when_reading_rods --minPruning 3

--maxNumHaplotypesInPopulation 200 -ERC GVCF --max_alternate_alleles 3 -contamination 0.0

-L Agilent_human_exome_v5_S04380110/S04380110_Covered.baits.nochr.w100.nr.bed

For each chromosome, we ran CombineGVCFs in batches of ~65 samples using parameters:

-T CombineGVCFs -R hs37d5.fa --disable_auto_index_creation_and_locking_when_reading_rods

--variant sample1.vcf.gz --variant sample2.vcf.gz ... -isr INTERSECTION

-L $chr -L Agilent_human_exome_v5_S04380110/S04380110_Covered.baits.nochr.w100.nr.bed

Then, for each chromosome, we ran GenotypeGVCFs on the output of CombineGVCFs using parameters:

-T GenotypeGVCFs -R hs37d5.fa --disable_auto_index_creation_and_locking_when_reading_rods

--variant group1.vcf.gz --variant group2.vcf.gz ... -isr INTERSECTION

-L $chr -L
Agilent_human_exome_v5_S04380110/S04380110_Covered.baits.nochr.w100.nr.bed

SAMtools/BCFtools: [samtools version: 1.1-30-g7f47a7c, bcftools version: 1.1-113-gd991f3f, htslib version: 1.1-104-g948a68c]:

All-site, all-sample BCF files from all the BAM files were generated using `samtools mpileup`. This processing was split into 100Mb chunks across the genome using parameters:

samtools mpileup -t DP,DPR,INFO/DPR -C50 -pm3 -F0.2 -d2000 -L500 \
-l
Agilent_human_exome_v5_S04380110/S04380110_Covered.baits.nochr.w100.nr.bed \
-g -r $chr:$from-$to -b $bam_list -f hs37d5.fa > $chr:$from-$to.bcf

For each chunk, variants were called using `bcftools call` using parameters:

bcftools call -vm -f GQ $chr:$from-$to.bcf | bgzip -c > $chr:$from-$to.vcf.gz

On chromosome X, male samples were treated as diploid in the pseudo-autosomal regions (X:60001-2699520 and X:154931044-155270560) and haploid otherwise using the `-X' option in `bcftools call`.

From the two initial callsets produced using GATK and samtools, we created a consensus call-set by intersecting the set of bi-allelic sites produced by each caller. The concordance between the two call-sets for SNPs was 95%, discordant genotypes were set to missing, variants with >1% missing genotypes were excluded, and all subsequent analyses performed on this consensus call-set. The ratio of transitions to transversions (Ts/Tv) in the consensus call-set was 2.52 for all variants and 2.42 for singletons (37.3% of SNPs), with much lower Ts/Tv ratios for discordant calls.

We did not attempt to identify large deletions (e.g. >100bp) in our dataset, as these methods remain inaccurate for medium depth sequencing (the ~40x depth used here for most samples); we note that as a consequence we will have called a hemizygous LOF genotype (with the other allele a deletion) as a homozygote, but that this would still lead to complete allelic inactivation.

Estimating variant quality

Estimation of error rates in exome sequencing data typically includes comparison with different technologies, or pedigree analysis, which whilst effective may miss systematic errors. A recent study(37) used a haploid human cell line, reporting heterozygous calls in the data as a proxy for error rate. Here, we use a similar strategy by examining the number of heterozygous calls within long autozygous stretches in our final call-set of 3222 Pakistani-heritage adults. We expect that long stretches of >10Mb autozygosity have occurred due to a recent inbreeding event, and therefore we can attribute heterozygous calls within such regions to either a de-novo mutation/gene conversion event, or a sequencing error. As the regions of autozygosity across these individuals occur due to random recombination events, we are able to measure errors occurring on multiple haplotypes and on joint called data.

We restricted autozygous regions considered to those at least 10Mb long and with no other region in that individual starting within 3Mb, and excluding the terminal 1Mb of each autozygous region. These stretches are long (often comprising a fifth of an entire

chromosome) and heterozygous calls within them are equally likely to be seen in the middle of such stretches as compared to the ends (t-test P=0.74). Overall we saw 88192 heterozygote calls in 12676437517bp of autozygous callable exome sequence, a rate of less than 1 per 100,000bp. Downsampled to the length of a single genome, this amounts to 19831 false heterozygotes per genome, within the range previously reported of 15000–30000 false heterozygotes per haploid genome sequenced at high coverage without PCR artifacts(38). This is still much higher than the expected rate due to new mutations, which is approximately 10^{-7} per bp (6 generations at a mutation rate of $1.5*10^{-8}$ per bp per generation), so most heterozygote calls within the autozygous sections are due to errors.

Final variant filtering

We next used the Variant Quality Score Recalibration (VQSR) tool within GATK to calibrate the probability of variant call error, and further filter the dataset based on this single estimate for the accuracy of each call. We trained the VariantRecalibrator Gaussian mixture model using a set of true sites from HapMap project variants(35). We used the following metrics to train the model - for SNPS: QD, FS, MQRankSum, ReadPosRankSum, BaseQRankSum, MQ, InbreedingCoeff, SOR, GQ_MEAN, NCC; and for indels QD, FS, ReadPosRankSum, BaseQRankSum, MQ, InbreedingCoeff, SOR, GQ_MEAN, NCC. After calibrating using the heterozygotes in autozygous regions identified above as false positives (**fig. S5**), we chose the VQSR 99% filtering threshold for SNPs, and no VQSR filtering on the indels (VQSR did not increase specificity for indels). We used this as the final dataset reported in the main results section of this paper. In this final dataset, the SNP heterozygous call error rate is 1.47% (the rate of false heterozygote calls in autozygous regions as a fraction of the total rate of heterozyote calls in non-autozygous regions) and the corresponding indel heterozygous call error rate was 1.63%. The ratio of transitions to transversions (Ts/Tv) in the final call-set was 2.56 for all variants and 2.50 for singletons.

As an additional method to estimate error rate, we used 176 pairs of known duplicate samples (independent blood samples taken from Born In Bradford mothers at separate pregnancies). By examining the replication rate of heterozygous calls within autozygous sections in these individuals, we can classify our calls into those that are concordant; those likely to be due to systematic reasons (such as read mis-alignment to the genome, or de novo mutations) and those that are discordant; which are likely to be due to random issues in the sequencing process or sampling of reads during variant calling. For SNPs, there were on average approximately 27500 homozygous alternate calls/individual of which 1650 were in autozygous regions, with 99.25% replication in duplicate samples (99.6% in autozygous regions). With VQSR at 99% we lost 1.3% of these 27500 calls. There was a mean of 20 false heterozygote SNP calls/individual, with 50.1% replication in duplicate samples. For indels, there were mean 2068 homozygous alternate calls per person of which mean 123 were in autozygous regions, with 91.8% replication in duplicate samples (92.9% in autozygous regions). There was a mean of 2.6 false heterozygote indel calls/individual, with 24.8% replication in duplicate samples (**Table S5**). Using the 176 pairs of duplicate DNA samples from individuals to estimate the reproducibility of homozygous alternate allele genotypes in the final call set, we found a 0.5% SNP homozygous genotype discordance rate (0.3% within autozygous

segments) and a 8.2% indel homozyous genotype discordance rate (7.1% in autozygous regions).

Functional annotation of variants from exome sequencing

Loss-of-function (LOF) annotation was performed using the Loss-Of-Function Transcript Effect Estimator (LOFTEE, version 0.2, available at https://github.com/konradjk/loftee) a plugin to the Ensembl Variant Effect Predictor (VEP, version 77) based on GENCODE version 19 for the GENCODE basic set(*39*). LOFTEE considers all stop-gained, splice-disrupting, and frameshift variants, and filters out many known false-positive modes, such as variants near the end of transcripts and in non-canonical splice sites, as described in the code documentation. As a variant may have multiple different effects on different transcripts, the annotation of function is based on the most severe consequence per variant in the order as defined in Ensembl (http://www.ensembl.org/info/genome/variation/predicted_data.html).

Identification of autozygous genomic segments

We applied a hidden Markov model (HMM) first utilized in(*40*) (https://samtools.github.io/bcftools/bcftools.html) to identify regions of homozygosity (absence of heterozygous variation) due to parental relatedness with the important addition of utilising the fine scale sex averaged human recombination map(*41*). The allele frequency information was obtained using all 3,222 exomes and the transition parameters between autozygous and non-autozygous sections were learnt from the data using a Viterbi training scheme (segmental k-means algorithm), with the initial probabilities of being in the N (non-homozygous) or H (homozygous) state at the start of each chromosome set to equal. The resulting state assignments given by the Viterbi sequence with the optimal parameters comprised our inferred homozygous and non-homozygous tracts. All regions of homozygosity identified were >10kb in length.

The command line used for the inference using bcftools roh was:

bcftools roh -G30 -a1e-8 -H1e-8 -e - -G30 -V -m genetic_map_chr{CHROM}_combined_b37.txt

We then compared our estimates of genetic autozygosity with those from self-stated estimates (**fig. S4A**). This confirmed our results in two ways. First, the estimate of genetic autozygosity corresponded with those from theoretical predictions from pedigree data, with offspring with higher parental relatedness having higher autozygosity. Second, we found that the median autozygosity from our genomic estimates was elevated over what we would expect in otherwise outbred individuals (first cousins, 6.25% and second cousins, 3.125%). This can be attributed to longstanding endogamy in the population which would lead to additional historic identity by descent. Additionally we noted that the variance even within individuals who stated that they were children of first cousins is extremely large, with estimates ranging from 0 to 25%.

After removing artifacts near locations close to the centromeres where we had low coverage in the sequencing data, we calculated the number of individuals that are autozygous at every site across the genome (**fig. S4B**), from which we can draw the following conclusions.

Every position in the genome contains at least one individual who is autozygous at a certain site, with a mean of 210.

The distribution of individuals who are autozygous at a site is not significantly different from random (Shapiro-Wilks test, see main text). The expectation was calculated by taking a weighted average of the overall inbreeding coefficient of each individual and assuming that the distribution of such sections would follow a binomial distribution and by approximation normal across 3,222 samples. Furthermore, the fact that these segments are randomly located across the genome and that such segments lie in hundreds of individuals with differing haplotypes means heterozygote sites can be used as a means of quality control of variants without any bias related to genomic location.

We find that 94.9% of all rhLOFs lie within the identified autozygous sections. This confirms that the overwhelming majority of homozygous genotypes were inherited from the parents at conception and hence not mosaics, ruling out mosaicism as a possible major reason for incomplete penetrance.

Exome Aggregation Consortium (ExAC) and Icelandic population comparisons

We compared genes containing rhLOF in our dataset versus the genes containing homozygous LOF (restricting to high confidence variants including PASS all filters, and 80% call rate across all samples) in the Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org, data release 0.3 with VEPv79 annotation, accessed June 2015). A total of 1775 of 26915 genes in ExAC from 60706 individuals contained homozygous LOF genotypes. We included variants of all allele frequencies, as ExAC comprises multiple diverse ethnicities.

We compared genes containing rhLOF in our dataset versus the 1,171 genes containing homozygous LOF or compound heterozygous LOF for variants with a minor allele frequency <2% in the Icelandic sequenced dataset, using both direct sequenced and imputed Icelandic population data from 104,220 samples.

Expected number of knockout variants to be seen in larger cohorts

The sequence data obtained from 3,222 healthy individuals provides us with a sample of the number and diversity of LOF variants (in both heterozygous and homozygous states) available in this specific population and we can use this to obtain estimates on the number of knockouts we expect to see in future sequencing up to 100,000 samples. In 3,222 individuals we observe 6,444 copies of each gene sampled from the population. In 100,000 people of first cousin offspring (each individual having 6.25% identity by descent), we expect to see 6250 homozygosed copies of each gene. Since this number is smaller than the number of observed copies (6,444) we could downsample the observed data set of haplotypes. As this does not account for heterozygous variation that would be incompatible with healthy life when homozygosed, we then reduce the estimated number of variant sites seen by 13.7% as this is the expected depletion of LOFs estimated using our subsampling approach. We plot the results as a function of sample size in **fig. S3**. We expect these to be conservative estimates because we are including the regions that are autozygous in these individuals in our initial sampling. Based on this current data we note that there are a large number of

genes yet undiscovered in which knockouts are likely to be compatible with adult human life, and that the discovery rate of these genes does not appear to plateau even if a study were to sequence 100,000 subjects with closely related parents. We also expect different sets of variation in different ethnicities and populations.

Selection signatures of LOF variants in autozygous and non-autozygous regions.

We examined direct selection on recessive deleterious LOF variants by comparing the rates at which we observe variant sites of different mutational classes within and outside of autozygous tracts within individuals. Since the individuals ascertained in this study were healthier adults when sampled, recessive LOF variants that lead to death as an embryo, foetus or child, or result in severe early-onset Mendelian disease should be less frequent in autozygous tracts, where they are exposed, while still present in non-autozygous portions of the genome. To control for the differing amount of autozygosity within the individuals as well as any differences in the relative rates of diversity inside and outside of autozygous sections, we normalize the total number of LOF genotypes seen in each section by the total number of variants in each section. We assess significance of the difference between relative rates of variants in the autozygous and non-autozygous portions of the genome using a t-test where the autozygous and non-autozygous rates are paired. To adjust for inconsistencies that might arise due to sampling error from individuals with small regions of autozygosity we further restrict our comparison to samples that have a total autozygous length of >=5%. We then examine the results of this statistic using different classes of variants (**fig. 2A**). We show as a control that rates of synonymous variants are not significantly reduced between the autozygous and non-autozygous sections, and then show a significant reduction in the rate at which we observe LOF variants within autozygous sections as compared to the non-autozygous sections, demonstrating direct selection against highly deleterious recessive variants.

Quantifying the depletion of LOF genotypes

The previous analysis on variant genotypes within and outside of autozygous tracts suggests that synonymous variants are an effective neutral control, and that there has been selection against homozygous LOF variation. In the allele frequency range under 1% there were 16,163 segregating LOF variants, with >= 1 non-reference homozygote genotype (rhLOF) found at 847 variants. We then matched the LOF variants to randomly selected synonymous variants with the same allele frequency and observed at how many of these variant sites non-reference homozygotes occurred, repeating the random selection process 10,000 times to estimate the distribution expected under neutrality. **fig. 2B** shows the resulting distributions and variant counts. We see a 13.7% deficit in variants containing rhLOF genotypes compared to the mean of the distribution for matched synonymous sites. This estimate makes no use of our earlier autozygosity assignment, so is not biased by any inaccuracy in autozygosity assignment. It is an average over the range of allele frequencies below 1% homozygosed in our sample.

The average number of recessive lethal variants carried by humans

The concept of lethal equivalents, defined as the expected number of heterozygous mutations in a single individual that would result in lethality when homozygosed, was first introduced in 1956, when it was estimated by a regression of the degrees of parental relatedness on the viabilities of their offspring(*42*). Whist methodologically sound, the estimate of the inbreeding coefficient of the infants was obtained theoretically by simply examining the known relationship of the parents. As we have seen, information obtained from recent pedigrees often does not capture the total amount of relatedness present in samples with extensive historic parental relatedness. Secondly, due to small sample sizes as well as the use of a theoretical inbreeding coefficient F, the results of this approach vary depending on the choice of the regression model used. Finally, the approach assumes that the socio-economic backgrounds as well as care received with regard to complications that might arise during pregnancy is the same across groups with different relatedness structures. From long standing cohort studies in the UK(*43*) we know that communities with more historic cousin marriage practices also have higher levels of health deprivation and education, which may have a significant impact on the birth outcome.

To minimize these limitations, we carried out a modern version of this approach by examining a dataset of 13776 mothers from the BiB cohort for which we had pregnancy outcomes. This dataset is particularly suited to this approach due to its large sample size from a single long-standing study in Bradford, UK. All of the mothers presented to the same maternity ward during their pregnancies and received standardised information on pre- and post-natal care. However the major difference from previous studies using this approach is that we had a direct DNA based estimate of the distribution of autozygosity (which we used in place of the the inbreeding coefficient) as a function of self-stated parental relatedness in this community from **fig. S4A**. We chose to remove pregnancies with the birth of more than one child to further remove bias.

We calculated the Survivability (S) as the fraction of pregnancies that resulted in a healthy offspring surviving to at least one year of human life and carried out a weighted least squares regression as first described by Morton et al.(*42*) to estimate the coefficients and standard errors of the A and B terms in their model. In **fig. 2C**, we report our estimates, as well as those obtained from other previously reported studies(*7*, *8*). The results obtained for A and B are 0.004970 (weighted least squares regression, SE 0.001135, P=0.02205) and 0.22711 (SE 0.03745, P=0.00902) respectively. So as to be conservative with our results using this approach across all datasets, we choose to report the estimate of the number of recessive variants incompatible with healthy life in a human genome as B±SE.

Along with these estimates based on epidemiological data, we also obtain a direct estimate of the number of recessive lethal variants based on the suppression of homozygote genotypes (as described above). For each allele frequency, we calculated the number of variants that are depleted compared to the neutral expectation of synonymous variants. We then take a weighted sum with the allele frequency to get the total number of LOF variants (not sites) that are expected to be incompatible with adult life across 3,222 individuals. Standard errors are computed by a block jack-knife across individuals. As these variants are found in all 3,222 individuals we take the average to obtain an estimate for the number of heterozygous variants carried by a single individual that

would be lethal or result in severe disease if homozygosed. In **fig. 2C**, we also include another recent direct estimate for a similar measure using recessive disease pedigrees(*44*).

Analysis of LOF load in different population cohorts and relation to demographic structure

We carried out a comparative analysis of variants at functional loci and examined the mutational burden of LOF variants as identified by annotation between populations. To obtain a comparative dataset to use for the analysis, we obtained calls for worldwide populations from the 1000 Genomes Project phase 3 and restricted analysis to the same exome bait regions as our dataset and ran our annotation pipeline using the same settings as described earlier. We then annotated ancestral and derived alleles using Ensembl Compara's 8 primates EPO alignment (http://www.ensembl.org/info/genome/compara/).

We calculated a statistic described in (*40*, *45*) which compares two populations, given a particular category of sites, in terms of the number of derived alleles found at sites within that category in one population rather than the other. The rationale behind the statistic is to compare the haploid load of mutations of a particular class in one population versus another. If the mutation rate per year is identical in both populations after the two populations have diverged and have undergone different demographic patterns, then the difference in the haploid mutational load has to have occurred due to selection. To aggregate information across multiple individuals, we use observed derived allele frequencies in each population and compute the statistic as follows. At each site *i* we write the observed derived allele frequency in population A as $f^A_i = d^A_i / n^A_i$, where n^A_i is the total number of alleles called there in population A and d^A_i is the number of derived alleles called. Similarly we define f^B_i in population B. Then if C is a particular category of protein-coding sites and *S* a set of synonymous sites, we define

$$L_{A,B}(C) = \frac{\sum_{i \in C} f_i^A (1 - f_i^B)}{\sum_{i \in C} f_i^B (1 - f_i^A)}$$

as a measure of the relative number of derived alleles found more often in population A compared to population B. We then define the ratio

$R_{A/B} = L_{A/B}(C) / L_{A/B}(S)$,

normalising by the value over putatively neutral synonymous sites, to mitigate any population-specific differences in overall mutation rate, as well as population-specific reference biases and any calling biases between our dataset and those from the 1000 Genomes. This is akin to the approaches in (*40*, *45*) where biases to do with branch shortening and deamination errors between ancient and modern genomes are mitigated. Estimates of the variance in $R_{A/B}$ were obtained using 100 block jackknifes on the set of sites in *C*.

Using our definition of LOF variants we then compared the value of $R_{A/B}$ in the 1000 Genomes Project cohort and our dataset to look for differences in historic selection for variants of this specific class. **Fig. 2D** shows $R_{A/B}$ for LOF variants found in one population as a comparison with that from an outbred European ancestry population,

CEU. In most populations there is no significant difference. However, in the Finnish population sample (FIN), which underwent a severe bottleneck, we see significant differences compared to many other populations (**Table S6**). Our population (BB in **fig. 2D**), which unlike FIN has a high heterozygosity so shows no evidence of a comparable bottleneck, shows a similar reduction in $R_{A/B}$ to FIN. We conclude that the reduction in load of severely deleterious mutations caused by homozygosing arising from endogamy in the BB population has been comparable to that arising from increased homozygosity during the bottleneck in the FIN population. We note that the CHS population from rural Southern China also has an increased $R_{A/B}$ value without a decrease in heterozygosity, indicating that it may also have been subject to historical endogamy. The relatively high value of $R_{A/B}$ for the GBR population may be a consequence of one third of its samples coming from Orkney, a small island archipelago north of Scotland with a small long term endogamous population.

Comparative genomics, human versus mouse

In order to understand the accuracy with which mouse models reflect human phenotype, we compared 215 genes with rhLOF in our dataset to mouse gene knockout data, requiring an exact 1:1 mouse:human gene ortholog. We removed genes with i) a one to many, or a many to one cross-species mapping; ii) removed genes where the human ortholog also has an OMIM annotation; iii) considered a mouse gene essential if in any strain or experiment reported in the Jackson Labs Mouse Genome Informatics Mammalian Phenotype database that there was a lethal phenotype observed. Of these, there were 52 genes where a lethal mouse phenotype had been reported on at least one genetic background. Properties of genes essential in mouse but not in humans showed no significant differences to those non-essential in both species across protein divergence (dN/dS), number of gene duplications in humans since species divergence, and gene expression (**fig. S6**) (all data downloaded from Ensembl BioMart).

Druggability and clinical approval analysis

We annotated genes with LOF variants with information concerning potential druggability – that is the potential for modulation of the protein target by a water-soluble small molecule drug. Druggable proteins usually contain a defined binding pocket or active site, which could act as a site of action (pharmacophore) for an orally bioavailable small molecule drug. We grouped proteins into four druggability classes, based on a collation of complementary published annotations of the potentially druggable genome and publically available databases of small molecules in the drug gene interaction database ((DGIDB); http://dgidb.genome.wustl.edu/; (v1.72). Targets in class D1 have a known drug recorded in dgidb; class D2 have small molecule tools recorded in CHEMBL (www.ebi.ac.uk/chembl) which may be in current development within pharmaceutical companies, and could be used as tools in animal and cellular models; class 3 are homologous to class 1 or class 2 targets described in several druggability publications collated in DGIDB; class 4 are predicted to contain a potentially druggable pharmacophore based on de novo structure-based druggability prediction using the dogsitescorer tool(*46*) (dogsite.zbh.uni-hamburg.de).

A protein can be considered a potential biopharmaceutical target if it is present in the cell membrane or extracellular space. Consequently we designated a protein as a biopharmaceutical target if it was reported to be extracellular or transmembrane in the Gene Ontology location category(*47*).

We compared the ultimate success or failure of drug development for targets contained within the different LOF gene datasets using a recently published dataset of drug development outcomes for 19,085 target-indication pairs(*11*)(citeline.com/products/pharmaprojects/). We used a chi-squared test to evaluate differences in the ultimate EU/US approval rate for targets with observed LOF variants, compared to background information target-indication pair approval.

Protein-protein interaction network analysis

Collections of genes with loss of function (LOF) and gain of function (GOF) variants (**table S4**) were compared against a genome wide background of molecular interactions derived from the STRING database (string-db.org/). Interactions were organised into seven categories, Binding, Reaction, Activation, Expression, Catalysis, Post Translational Modification and All Interactions. The distribution of interactions were observed to be non-normal in most cases, probably due to missing data. We therefore compared the distributions of interactions between gene collections using a non-parametric Kruskal-Wallis test to obtain a p-value (R, http://www.r-project.org/). Low medians were seen across several of the seven interaction groups, therefore 5% and 95% quantiles were also reported in addition to median values for each group.

Sanger sequencing validation

Previous work suggests that LOF variants might be enriched for sequencing errors(*1*). We recalled 19 subjects for validation: 12 because of a rhLOF genotype in a highly expressed blood gene (for use in RNA expression in blood studies), and 7 with diverse genes of potential interest. Of the 37 homozygous LOF genotypes in these subjects, we validated 35 rhLOF genotypes (2 could not be assayed) using a different method (Sanger dideoxy sequencing) in these independent samples (**table S1**).

Sanger sequencing was performed on PCR products using an Applied Biosystems (Waltham, USA) 3730xl DNA analyser and big dye terminator 3.1 cycle chemistry. We sequenced all samples with rare variant allele genotypes, and a control sample, for the sites selected.

Western blot and RNA expression validation

Selected subjects were recalled for a further blood sample. Whole blood was preserved for RNA immediately upon venesection using PAXgene system (Qiagen, Venio, The Netherlands). Heparinized whole blood for protein assays was stored/transported at room temperature overnight and peripheral blood mononuclear cells were isolated by density gradient centrifugation with Lymphoprep (Stemcell Technologies, Vancouver, Canada).

Peripheral blood mononuclear cells were lysed using CelLytic M plus protease inhibitor (Sigma-Aldrich, St. Louis, USA). Cell extracts were quantified by BCA (Pierce, Waltham, USA) and stored at -80°C till required. 2-10ug protein/lane were run on either 4-12% Tris Glycine or 10-20% Tricine Novex pre-cast gels (Life Technologies, Waltham, USA). Gel type used was determined by predicted protein size, specifically DPYD Tris-Gly gel 2ug; GCA Tricine gel 7.5ug; LSP1 Tris-Gly gel 3ug; SAMD9 Tris-Gly gel 10ug and MSRA Tricine gel 10ug. Gel electrophoresis and transfer to Novex 0.45μm PVDF membranes (Life Technologies) were done using the XCell surelock system (Life Technologies). Membranes were incubated with antibody at 4°C overnight before development using Fast Western SuperSignal West Pico kit (Pierce) and imaged using the hyperprocessor (Amersham Pharmacia Biotech, Little Chalfont, UK) with CL-Xposure Film (Life Technologies). Membranes were stripped with Restore (ThermoScientific, Waltham USA) and reprobed with antibody against α-tubulin (α-tub, Abcam, Cambridge, UK) as loading control.

RNA was extracted using PAXgene Blood RNA kit (Qiagen), quantified by NanoDrop 8000 UV-Vis Spectrophotometer (Thermo-Scientific) and Bioanalyser (Agilent). 250ng total RNA per sample was labelled with the Illumina total prep RNA amplification kit (Ambion, Waltham, USA), and 750ng of labelled sample then hybridised to Illumina HumanHT-12v4 Expression BeadChip according to manufacturers instructions. Quantile-quantile normalisation and analysis was performed using Illumina GenomeStudio.

For a subset of genes known to be expressed in blood we observed the absence of protein on western blots using whole blood samples for 5 rhLOF genotypes (*LSP1*, *DPYD*, *GCA*, *SAMD9*, *MSRA*), weak protein (*EMR2*); and/or very low RNA expression compared to other samples for 6 rhLOF genotypes (*LSP1*, *SLC27A3*, *GCA*, *SAMD9*, *MSRA*, *EMR2*)(**fig. S1**). Extensive validation of LOF variants has been described elsewhere using RNA sequencing(*3*, *48*).

Manual annotation of rhLOF variants.

For each genotype we manually reviewed final bam sequences in the IGV viewer 2.3.59 (short read sequence data) for the rhLOF individual of interest, and 2 or more control subjects. We specifically looked for other nearby indels that would restore reading frame for rhLOF frameshift indels, and for adjacent variants that would eliminate stop codons, in addition to suspicious read alignment anomalies. We then examined the variant position in the UCSC genome browser, and in the ExAC browser, to look for potential issues with transcript and/or exon annotation that were missed by automated annotation with the LOFTEE tool. We reviewed rhLOF variants in annotated but non-conserved, alternative reading frames than the canonical exon. Splice variants in the last exon or 3'UTR, as well as ones with a nearby frame-restoring potential splice site, were also considered as suspect on the grounds that they were likely not to impact protein function. Variants thought highly unlikely to cause LOF were removed and listed with reasons in **Table S2**.

Comparison with a Mendelian disease database (OMIM)

We downloaded the OMIM morbidmap (http://www.omim.org/downloads on 12 May 2015), excluded records with "?" (unconfirmed or possibly spurious mappings) or "{}" (susceptibility to multifactorial diseases or infection) or "[]" (non-diseases that lead to apparently abnormal laboratory test values) annotation, included only records with "(3)" (molecular basis of the disorder is known) annotation and compared to our rhLOF gene list. We next selected only those genes with reported autosomal recessive inheritance (including those with multiple inheritance patterns) in OMIM.

We note that our study has an ascertainment difference compared to much of the Mendelian disease literature in that subjects were recruited as relatively healthy adults rather than patients with severe and young-onset diseases (and often from multiply affected families). We also recognize that health record analysis differs from phenotyping with prior knowledge of genotype.

Comments (columns in **Table S3**) were provided by independent review of the published phenotype-genotype associations and genome annotations by three clinical geneticists (AODL, ERM, RT), a clinician (DAvH) and a rare disease researcher (DMcA). Of the 38 individuals in **Table S3** with rhLOF genotypes in OMIM recessive genetic diseases as identified above, 6 have compatible entries and 3 are partially compatible. Of the remaining 29 individuals, 12 have comments on the genotype-phenotype association, and 8 have comments on the genomic annotation, with 1 having comments on both. Our review did not identify any possible reasons for absence of genotype-phenotype association for 9 individuals.

Primary healthcare records of Born In Bradford (BiB) subjects

All citizens of England are offered primary healthcare that is free at the point of use at general practices. Most practices have long used electronic health record (EHR) systems for the recording of diagnoses, symptoms, signs, according to the hierarchical system devised by Read (which maps to the internationally used SNOMED-CT) as well as prescriptions and test results. Structured primary care EHRs were obtained (and linked) for BiB participants registered with General Practitioner (GP) surgeries that use the TPP SystmOne platform. SystmOne has 100% coverage in Bradford and high coverage in surrounding areas. Records were extracted when the national unique health identifier (NHS number), surname, date of birth and gender were an exact match in SystmOne. From the full BiB cohort of 12450 mothers, 12333 (99.1%) were matched to their primary care records. Records were obtained from 18 months prior to study recruitment (Sep 2005 to Jun 2009), until end of November 2014, or until the participant died or withdrew from the cohort study if sooner. Total loss due to deaths, patients transferring to non-SystmOne practices, and withdrawing from the study, amounts to 0.4 years per person, with records available for 6.9 years per person, from a possible maximum 7.3 years per person. Of 2162 exome sequenced individuals available for analysis, 2145 (99.2%) had matched GP records.

For selected individuals with LOF variants of interest we also obtained lifetime electronic health records from SystmOne, and manually correlated Read codes with reported genotype-clinical phenotype associations in OMIM.

Primary healthcare record prescription- and consultation-rate analysis

Using the primary healthcare records of BiB subjects, we performed a prescription rate analysis, assessing all classes of prescribed medicines, irrespective of indication. Drug prescriptions are captured in the primary care extract using 10-digit British National Formulary (BNF) IDs. Published data show a count of unique BNF chapter headings in a patient's record can predict mortality and consultation rate(*17*). 10-digit BNF ID was truncated to a 4-digit chapter heading (the first two digits describe the organ system (e.g. cardiovascular), the second 2 digits the drug class (e.g. angiotensin converting enzyme inhibitors)), then distinct values per person were counted, so that multiple drug issues from the same BNF drug class were only counted once.

We calculated Consultation rate as clinical consultations per person year, using the formula 365 * (NClinicalConsultationDays / DaysInSystem). A consultation event in SystmOne is an abstract event logged by the application when data is entered by a user. To avoid over-counting due to use of multiple SystmOne client sessions in one patient visit, days on which at least one consultation event occurred were counted (NClinicalConsultationDays), with the following criteria: Include consultation events entered in a general practice setting; Include where the staff member role is GP/doctor or nurse; Include where consultation type is "Clinical"; Include where consultation method indicates face to face, telephone, home visit or unrecorded, but exclude others. The number of days per person transmitting GP data was derived from the SystmOne patient registration history. The extract begins 18 months prior to participant recruitment to the study. It ends (a) at end of November 2014, or (b) when the person withdraws from the study or (c) at death, whichever is earliest. A person can also transfer out of the system by registering with a non-SystmOne practice. These transfers out and back in were obtained from the SystmOne patient registration history, and the total number of days in the system per person was counted (DaysInSystem).

DaysInSystem was also used as a covariate in adjusted models, except where consultation rate was also in the model, as consultation rate already includes this component. Patient age is also used as a covariate in adjusted models, computed as age in years at end of last SystmOne GP registration period. Deprivation has been shown to predict GP consultation rate(*49*). Conventional measures of deprivation may lack validity across ethnic groups due to cultural differences in economic priorities and opportunities(*50*), whereas education has been shown to be effective in capturing variation in socio-economic position (SEP) across UK ethnic groups. The covariate education, with international qualifications equivalised based on UK NARIC [http://ecctis.co.uk/naric/ accessed 13 Feb 2015], was included in adjusted models as a marker for deprivation.

The effect of LOF genotypes on prescribing and consultation rate was examined by logistic regression using Stata v13. Patient age, education and days in the system were covariates for adjustment for all analyses. For consultation rate analysis, we also used a measure of mother's education level as the best BiB individual-level marker for deprivation.

A small proportion of participants were in the system and providing data for a very short period that may not be sufficient to detect differences in disease burden. By way of sensitivity analysis, the above analyses were repeated only including patients who provided >5 person years of GP data (n=2020). Some authors have noted that individuals

at the same level of education are not necessarily comparable in terms of socioeconomic (SES), and that multiple SES indicators should be used(*51*). Previous work in the BiB cohort has shown that SES variation within Pakistani ethnicity participants is captured by education, receipt of means tested benefits, material deprivation, subjective poverty and employment status(*52*). A third set of sensitivity analyses included these multiple SES indicators. None of the sensitivity analyses described here indicated conclusions that differed from the main analyses.

In addition to the primary rhLOF versus no rhLOF analysis (main text), we performed additional analyses in selected subgroups. We did not find association with prescription rate (logistic regression OR 1.006, 95% CI 0.966 - 1.045) or consultation rate (OR 1.029, 95% CI 0.964 - 1.090) in individuals (n=53) with rhLOF in OMIM confirmed recessive disease genes compared to individuals without rhLOF; nor on prescription rate (OR 1.005, 95% CI 0.968 - 1.043) or consultation rate (OR 1.038, 95% CI 0.978 - 1.094) in individuals (n=58) with rhLOF in orthologs of mouse knockout lethal genes. Findings were unchanged in additional adjusted analyses including age, education, duration of available data, and autozygosity.

Analysis of a subject with a predicted *PRDM9* knockout

The mechanism behind the localization and regulation of meiotic recombination in humans and other mammals are of considerable interest. Much research has been focussed on understanding the role of a single rapidly evolving gene, PR domain-containing 9 (*PRDM9*) in the molecular control of the distribution of meiotic double stranded breaks (DSBs) in mammals(*53*). Through the use of population based genome-wide analyses, bulk sperm sequencing as well as genome editing in model organisms, meiotic DSB sites have been characterised at high resolution.

Recent efforts have focused on understanding the patterns of recombination in species lacking PRDM9(*54*, *55*) as well as mouse *Prdm9* knockouts(*22*). These results suggest that in the absence of PRDM9 to localize breakpoints, most recombination is initiated at promoters and at other sites of PRDM9-independent H3K4 trimethylation. In humans, there has been extensive evidence to suggest that PRDM9 initiates and localises DSBs(*56*, *57*) and is a major determinant of hotspots. However, insights into its essentiality as well as direct functional studies in-vivo in humans have not been carried out. Here, we identified as part of our study an individual containing a homozygous knockout in the *PRDM9* gene, within a long autozygous region. We also studied one of three children (others in the family did not consent to research). We performed further experiments to define the recombination landscape in this family, and compared them to control data from a mother/child duo NA12878/NA12882 from a heavily studied three generation CEPH pedigree. NA12878 is homozygous for *PRDM9-A*, the most common allele of *PRDM9.*

Validation of genotype and functional validation

We examined a pileup of the exome sequencing reads for the variant as well as standard Sanger dideoxy-sequencing validation of the variant to confirm the quality of the genotype call (**fig. S7A,B**). We also observed 10 additional (not closely related) individuals in our study population to be heterozygous at this variant. Heterozygote

variants (but no non-reference homozygotes) at this genotype were also observed in ExAC (http://exac.broadinstitute.org/variant/5-23524525-C-T). No homozygotes for any other *PRDM9* LOF variants were observed in ExAC. We carried out careful manual annotation of the variant to ensure that the variant would result in a homozygous knockout of the gene, as described earlier.

Cell line studies

No primary cell lines or immortalised human cell lines stably expressing *PRDM9* were available, as *PRDM9* is specifically expressed only at meiosis. We therefore performed site directed mutagenesis to generate the chr5:23524525 T allele using the QuikChange II mutagenesis kit (Agilent Technologies) and full-length *PRDM9* cDNA cloned into the pCEP4 expression plasmid as described(*20*), and confirmed the expected allele and insert by sequencing. We transfected the plasmid into HEK293 cell lines as described(*20*), and assayed protein by western blot using anti-FLAG M2 (Sigma-Aldrich), anti-H3 (EMD Millipore (Billerica, USA), 06-755), and anti-H3K4me3 (EMD Millipore, 07-473). Chromatin Immunoprecipitation and qPCR at 5A and 22A hotspots in HEK293 cells was performed as described using anti-H3K4me3(*20*).

Illumina HiSeq X Ten whole genome sequencing

Whole genome sequencing data for the *PRDM9* minus duo was carried out by extracting DNA as described earlier. A single library (650 base pair inserts) was constructed for each sample. The libraries were multiplexed and sequenced across several lanes on the Illumina HiSeq X platform (paired-end sequencing, 151 cycles). Variant calling and filtering were carried out using the approaches described earlier.

10XGenomics long range Gemcode molecular phasing

The 10X Genomics (Pleasanton, USA) GemCode reagent delivery system partitions long DNA molecules (including DNA >100 kb) and prepares sequencing libraries in parallel across the partitions such that all fragments produced within a partition share a common barcode. A simple workflow combines large partition numbers with a massively diverse barcode library to generate >100,000 barcode containing partitions. Libraries on the mother and child were sequenced on an Illumina HiSeq 2500 instrument in Rapid Run mode. The GemCode Long Ranger Software maps short read data to original long molecules using the barcodes provided by the reagent delivery system, thus providing long range phasing information.

The DNA samples available for the PRDM9 duo, whilst of adequately high molecular weight for most genomics applications, were more fragmented than for the Coriell (Camden, USA) reference samples NA12878 and NA12882 (**Table S7**). Nonetheless, >90% of SNPs could be phased in each of the four samples. Unlike, e.g. detection of structural variants, identification of recombination breakpoints is less sensitive to the length of phased segments.

Obtaining recombination breakpoints using phased parent-child duos

We identify recombination crossover sites in the maternal meiosis leading to the child's genome using heterozygous sites in the mother and child that are informative about the transmitted chromosome. Of the possible allelic combinations of the mother

and child at each site (**Table S8**) four combinations are informative if only the mother is phased while eight are informative if both the mother and child are phased. To avoid false positives due to genotype and phasing errors, we used this information in a series of steps.

Initially we only used phase information from the mother and considered bi-allelic SNPs that were of high quality (10XGenomics phase quality scores at maximum). Once we located candidate intervals we removed crossover intervals that lay within 300kb of each other. Then we carried out manual visualisation of the crossover locations in the 10XGenomics Loupe browser to ensure that there were at least two informative SNPs on each side of a crossover that were of maximum quality. We then refined the crossover intervals using the child's phasing where it was available and overlapped with the phase set of the mother (this was in >90% of cases in NA12878 duo and ~70% in the PRDM9 duo), as well as considering sites where the phase quality was lower but was nevertheless consistent with the phased genotypes in the other sample of the duo. This approach resulted in 42 crossovers in NA12878 / NA12882 duo with a mean crossover interval length of 16.2kb, and 37 crossovers in the PRDM9 duo with a mean crossover length of 51.8 kb.

As a validation of our method, we compared the crossovers inferred in the NA12878 duo to a gold standard set derived independently for the same meiosis from segregation data across the complete three generation CEPH pedigree, all of whose members had been previously sequenced (*58*). This approach yielded a set of 53 crossovers with a mean interval length of 40.5kb. 40/42 (95%) of the crossovers inferred from the 10X Genomics phasing data overlap those from the pedigree based gold standard, signifying that our strategy identifies crossovers in parent-child duos with high specificity.

Statistical analysis to infer recombination landscapes

In order to compare the recombination landscape of the two samples, we utilized a method that has been previously described in several studies that have compared crossovers obtained from pedigree based studies with those from the fine scale recombination map(*18*, *59*). Here, we seek to examine the PRDM9 hotspot usage phenotype by comparing the crossover events that we call using long range molecular phasing with those obtained from high-resolution maps of meiotic DSBs in individual human genomes(*25*). These are tightly coupled with crossover locations observed from population sequencing and linkage disequilibrium (LD) based methods but provide information on the localisation of DNA binding by PRDM9 at a much finer scale. For additional confirmation and to compare our results with those from these previous studies we also utilized 32,996 autosomal hotspot locations inferred from genome-wide Phase II HapMap LD data(*57*).

The method estimates the proportion of crossovers (α) that occur in a set of predefined feature intervals. Care must be taken because we do not know the precise location of the crossover, only an interval (possibly large) within which it took place. We utilise the previously published method almost exactly as referenced above but describe it again here due to slight changes in implementation. We begin by obtaining the probability that crossover interval r overlaps a feature:

P(r overlaps a feature) = α + (1-α)P(r overlaps a feature by chance)

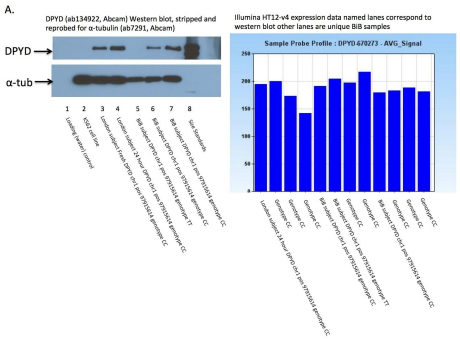
We estimated P(r overlaps a feature by chance) by randomly shifting the crossover r a normally-distributed distance (mean 0, std. deviation 200kb) 100 times and counting the fraction of these moves that result in the crossover r overlapping a feature (note that this probability differs for each crossover, due to the size of the interval containing r and the density of features in the local region). The likelihood of α for a crossover r is δ_r P(r overlaps a feature) + (1-δ_r)(1-P(r overlaps a feature)), where δ_r is an indicator variable that is 1 if r overlaps a feature and 0 if not. The likelihood of α for the full set of crossovers is the product of the per-crossover likelihoods.

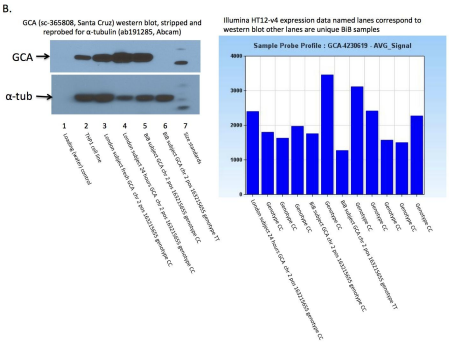
As in the previous references, the likelihood of α for all crossovers localised to an interval smaller than 30kb was calculated over a range of values between 0 and 1 in 100 steps and implemented using the NLM package in r. The 95% confidence interval of α was considered to include all values of α for which the log likelihood was within 2 units (the asymptotic cutoff) of the maximum log likelihood. Note that this estimation procedure naturally accounts for uncertainty in the location of LD hotspots and in the location of crossovers, as the overlap by chance is influenced by both the width of the estimated hotspots and the interval sizes in which we infer crossovers.

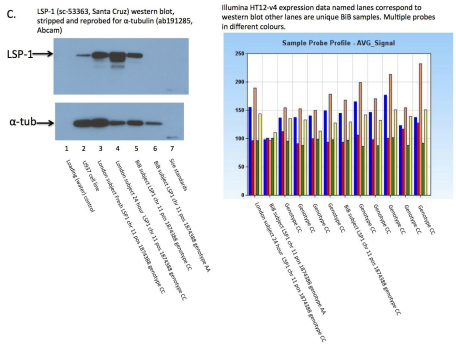
We report in the main text point numbers for the 23 crossovers in the *PRDM9-* duo and 34 crossovers in the NA12878 duo localised to within 30kb. When using all 37 (respectively 42) crossovers we get consistent results: for the PRDM9 A Union DSB hotspots: NA12878 duo - 55.2% [2 log unit confidence interval 38%-71%], *PRDM9-* duo - 9.7% [0%-26%]; and for the LD based hotspots NA12878 - 72.1% [54%-86%], *PRDM9-* - 18.3% [2%-34%].

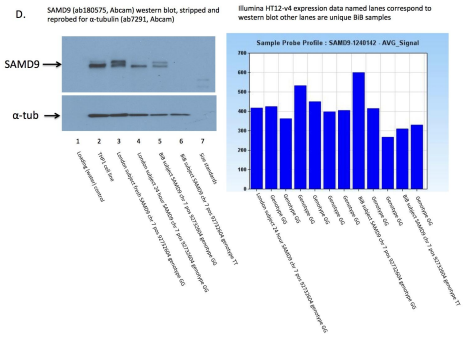
Examination of crossover overlap with GC rich regions, promoters or H3k4me3 sites

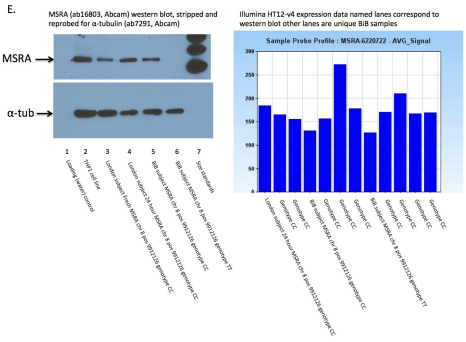
We obtained GC content, gene promoter and promoter flanking region information from Ensembl(*39*), and consolidated imputed data for narrow contiguous regions of enrichment (peaks) for H3K4Me3 ChIP-seq data for multiple cell types from the epigenomics roadmap project(*60*). Using the same approach as for overlap of our regions with PRDM9 hotspots we found no significant differences between the proportion of crossovers overlapping H3K4Me3 sites (NA12878 duo 4% [2 log unit confidence interval 0-8%], PRDM9 duo 5% [0-8%]), gene promoters (NA12878 duo 0% [0-4%], PRDM9 duo 0%[0-4%]) and their flanking regions, or in the GC content of 200kb of DNA (t-test P=0.64) from the middle position of each crossover interval.

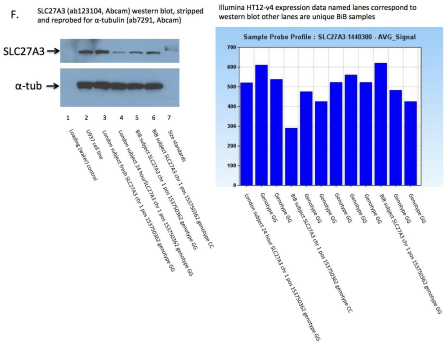

A.
DPYD (ab134922, Abcam) Western blot, stripped and reprobed for α-tubulin (ab7291, Abcam)
Illumina HT12-v4 expression data named lanes correspond to western blot other lanes are unique BiB samples
DPYD
α-tub
Sample Probe Profile : DPYD 6702713 : AVG_Signal


B.
GCA (sc-365808, Santa Cruz) western blot, stripped and reprobed for α-tubulin (ab191285, Abcam)
Illumina HT12-v4 expression data named lanes correspond to western blot other lanes are unique BiB samples
GCA
α-tub
Sample Probe Profile : GCA 4230619 : AVG_Signal


C.
LSP-1 (sc-53363, Santa Cruz) western blot, stripped and reprobed for α-tubulin (ab191285, Abcam)
Illumina HT12-v4 expression data named lanes correspond to western blot other lanes are unique BiB samples. Multiple probes in different colours.
LSP-1
α-tub
Sample Probe Profile : AVG_Signal


D.
SAMD9 (ab180575, Abcam) western blot, stripped and reprobed for α-tubulin (ab7291, Abcam)
Illumina HT12-v4 expression data named lanes correspond to western blot other lanes are unique BiB samples
SAMD9
α-tub
Sample Probe Profile : SAMD9 1240142 : AVG_Signal


E.
MSRA (ab16803, Abcam) western blot, stripped and reprobed for α-tubulin (ab7291, Abcam)
Illumina HT12-v4 expression data named lanes correspond to western blot other lanes are unique BiB samples
MSRA
α-tub
Sample Probe Profile : MSRA 6220722 : AVG_Signal


F.
SLC27A3 (ab123104, Abcam) western blot, stripped and reprobed for α-tubulin (ab7291, Abcam)
Illumina HT12-v4 expression data named lanes correspond to western blot other lanes are unique BiB samples
SLC27A3
α-tub
Sample Probe Profile : SLC27A3 1440300 : AVG_Signal

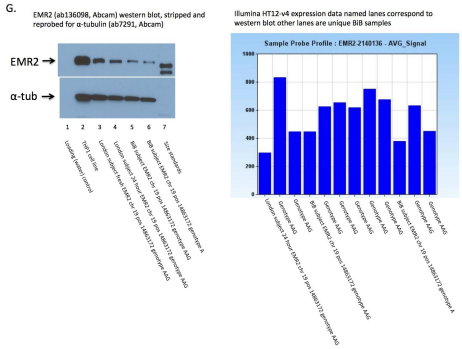

G.
EMR2 (ab136098, Abcam) western blot, stripped and reprobed for α-tubulin (ab7291, Abcam)
Illumina HT12-v4 expression data named lanes correspond to western blot other lanes are unique BiB samples
EMR2
α-tub
Sample Probe Profile : EMR2 2140136 : AVG_Signal

Fig. S1.

RNA and protein level validation of predicted LOF genotypes. Left panels show protein expression from Western blot. Right panels show probe-level Illumina HT-12v4 RNA expression array data. Panels A,B,C,D,E show absent protein in the LOF sample as assessed by Western blot. Panel G shows low protein in the LOF sample. Panel F shows apparently normal protein (presuming the antibody is specific) but low RNA levels in the LOF sample.

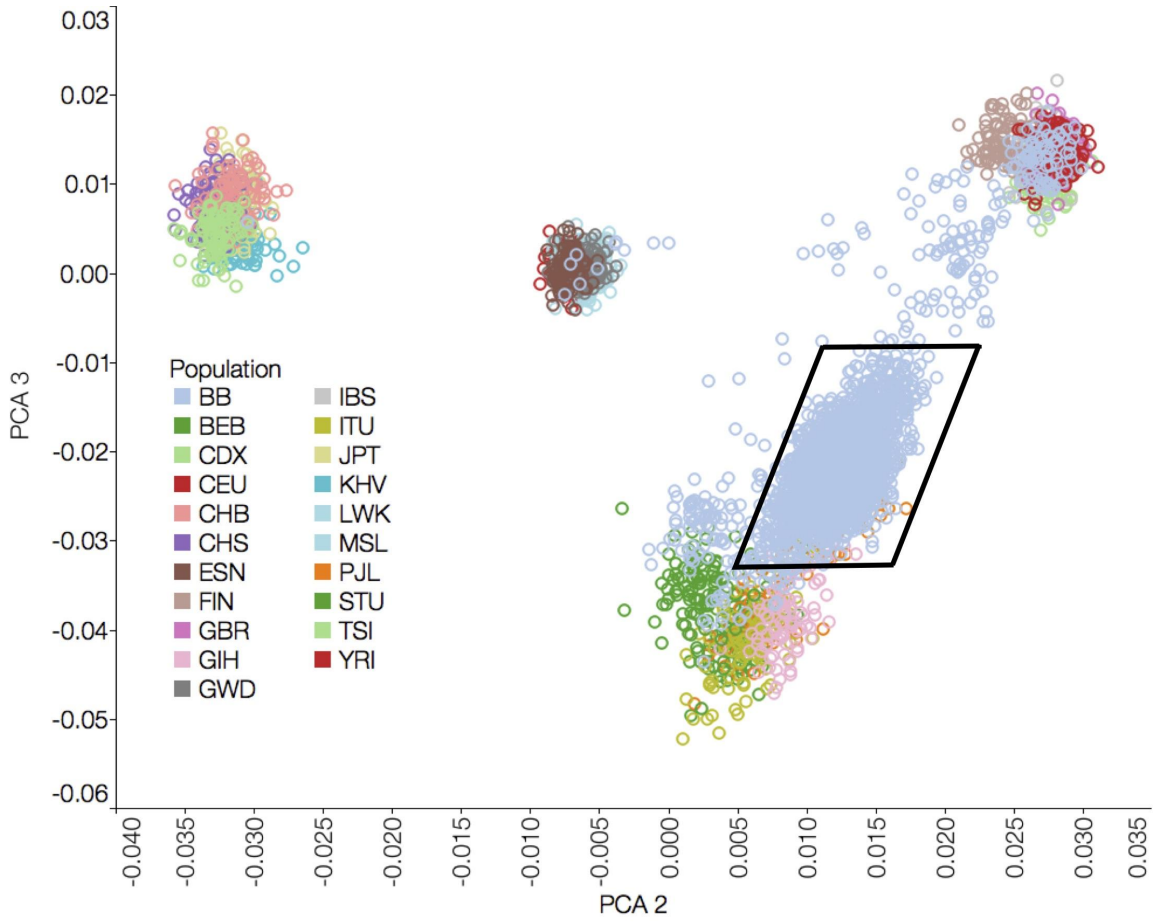


Fig. S2.

DNA-determined heritage. Principal component analysis was performed on individuals from Phase 3 of the 1000 Genomes Project, with the current dataset mapped onto this reference. We defined a polygon to include samples that were clustered together with the 1000 Genomes Project Pakistani population. Legends: BB - Birmingham and Born In Bradford; others - 1000 Genomes Project ethnic groups. PCA polygon (in black) refers to the subset of samples retained in the BB dataset for further analysis.

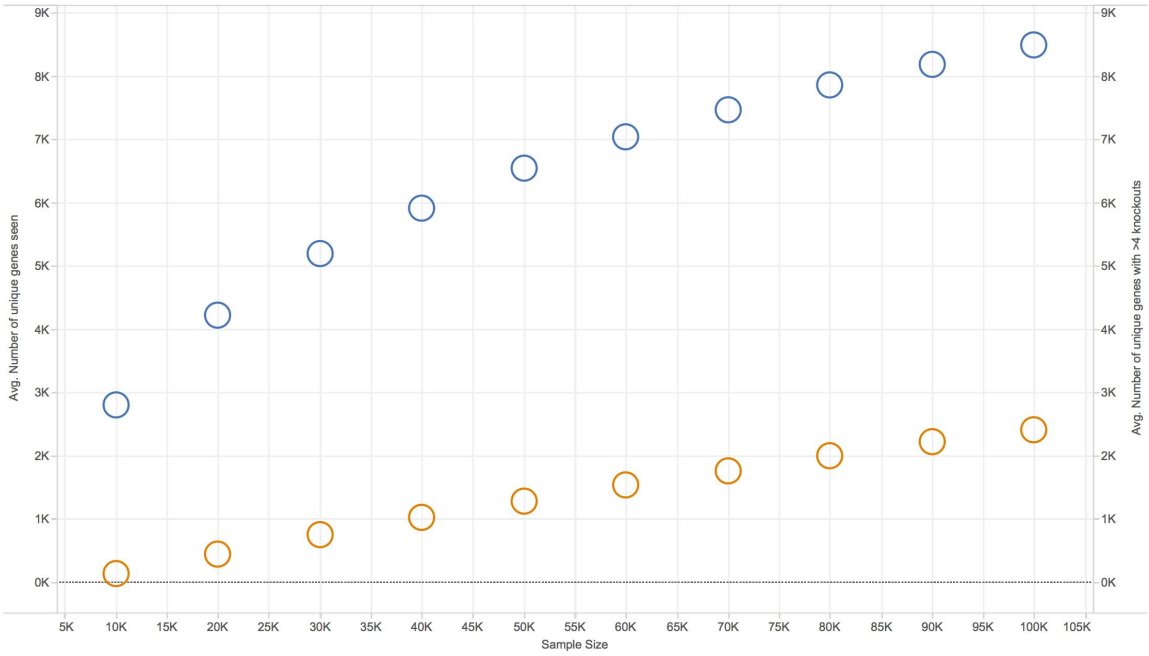


Fig. S3.

Estimates of number of rare homozygous knockout genes seen in larger cohorts (blue circles), and number of rare homozygous knockout genes with more than 4 individuals (orange circles).

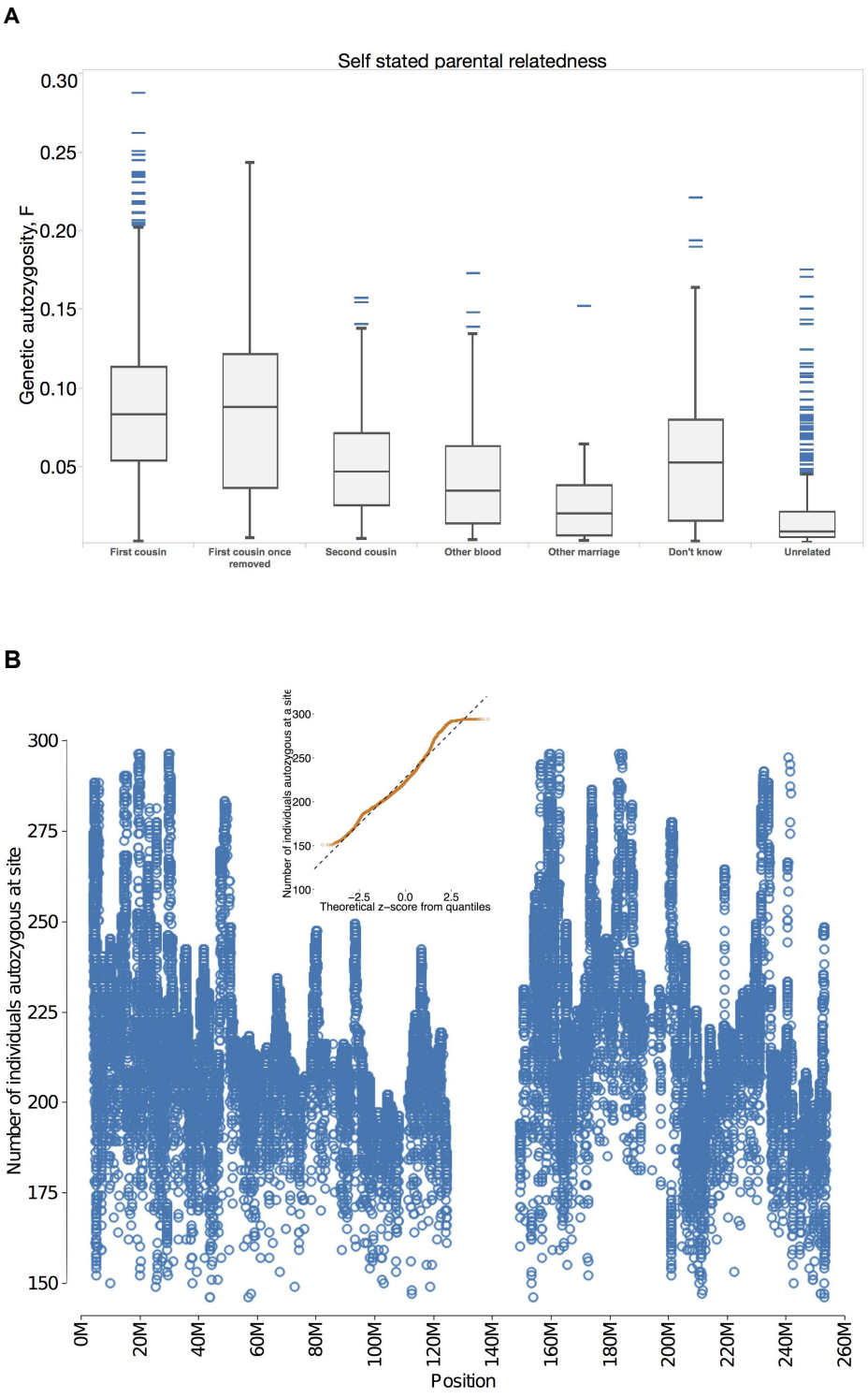


Fig. S4.

Analysis of autozygosity in populations. (A) Self-stated vs genetic autozygosity for individuals from Born In Bradford. Distribution of total length of genome in autozygous

stretches shown in boxplots, with individuals for whom no data was available indicated in the ‘Don’t know’ column. Genetic autozygosity, declines with self-stated parental relatedness and reflects theoretical expectations. (B) Levels of autozygosity observed across the genome. Number of individuals autozygous at a site in the protein coding sections of the genome (blue circles) across positions on chromosome 1 (shown as a representative example) on the x-axis. Inset in orange circles is a Q-Q plot of the theoretical expectation of the distribution, and shows normality of the data.

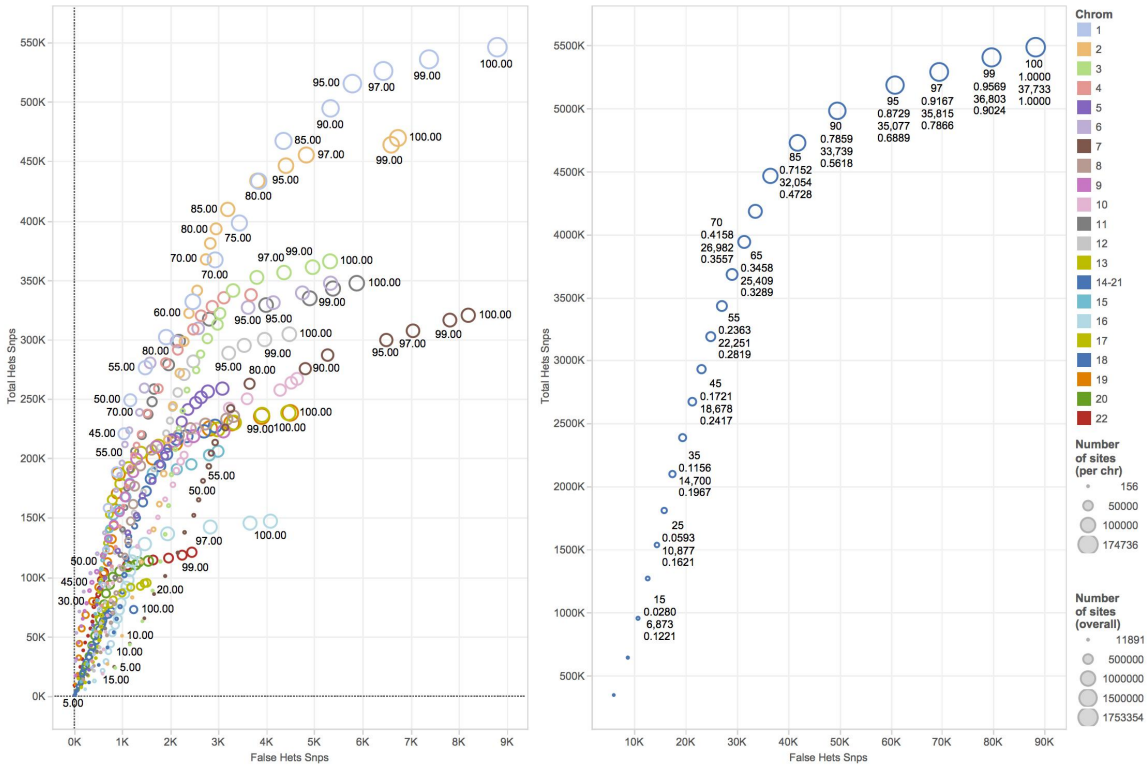


Fig. S5.

Using heterozygous calls within an autozygous section as a measure of variant quality control. Total number of discovered heterozygous sites in 3,222 individuals as a function of sites that were heterozygous within an autozygous block (x-axis) across individual chromosomes (colored circles on left panel) and summed across the genome (blue circles on right panel). Total number of sites overall is represented by the size of circle. Labels on each point represent in order: the VQSR false positive threshold reflecting the percentage of true positive sites left; (right panel only) the fraction of total sites left at that threshold, the mean number of hets per person and the number of false heterozygote sites left.

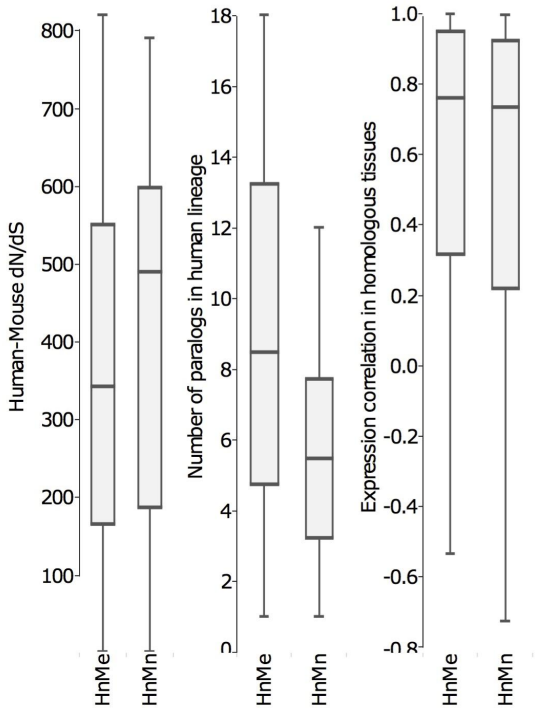
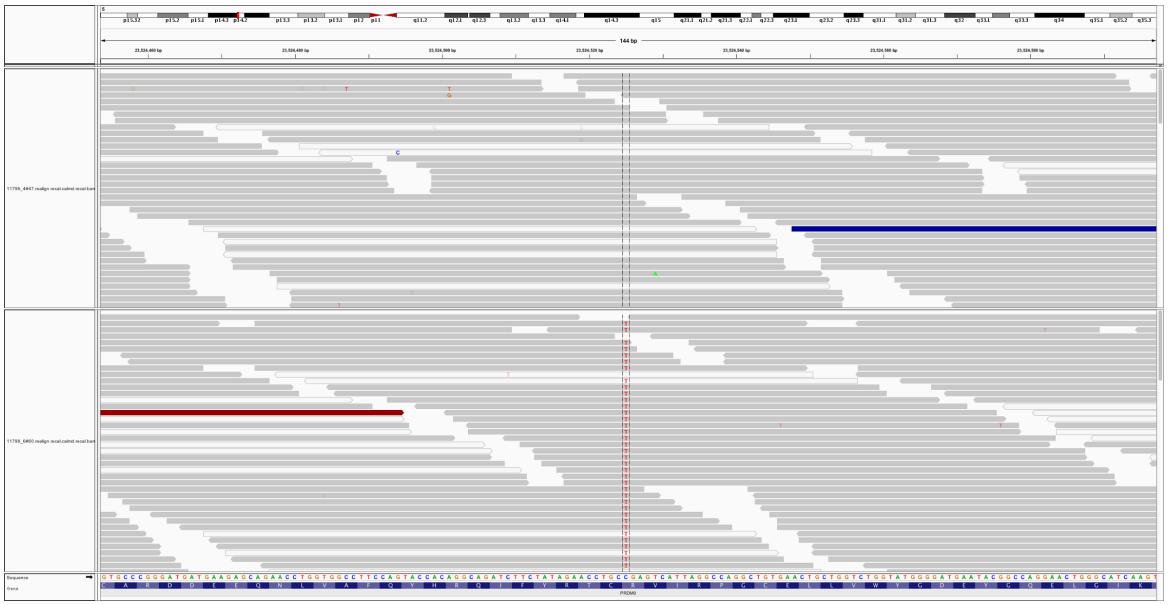


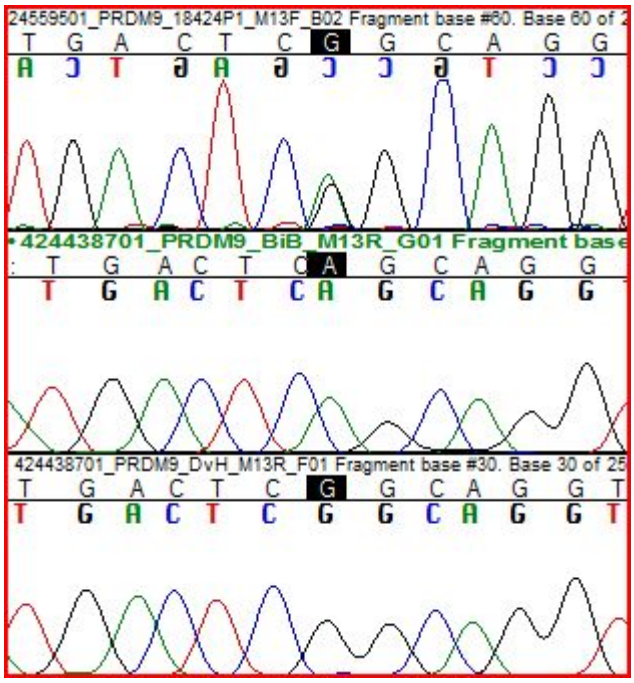
Fig. S6.

Mouse Human Orthologs. Properties of genes essential in mouse but not in humans (HnMe), compared with those non-essential in both species (HnMn) across protein sequence divergence (dN/dS), number of gene duplications in humans since species divergence and gene expression show no significant differences.

A



B


24559501_PRDM9_18424P1_M13F_B02 Fragment base #60. Base 60 of 2
T G A C T C G G C A G G
424438701_PRDM9_BiB_M13R_G01 Fragment base
T G A C T C A G C A G G
424438701_PRDM9_DvH_M13R_F01 Fragment base #30. Base 30 of 25
T G A C T C G G C A G G T

C

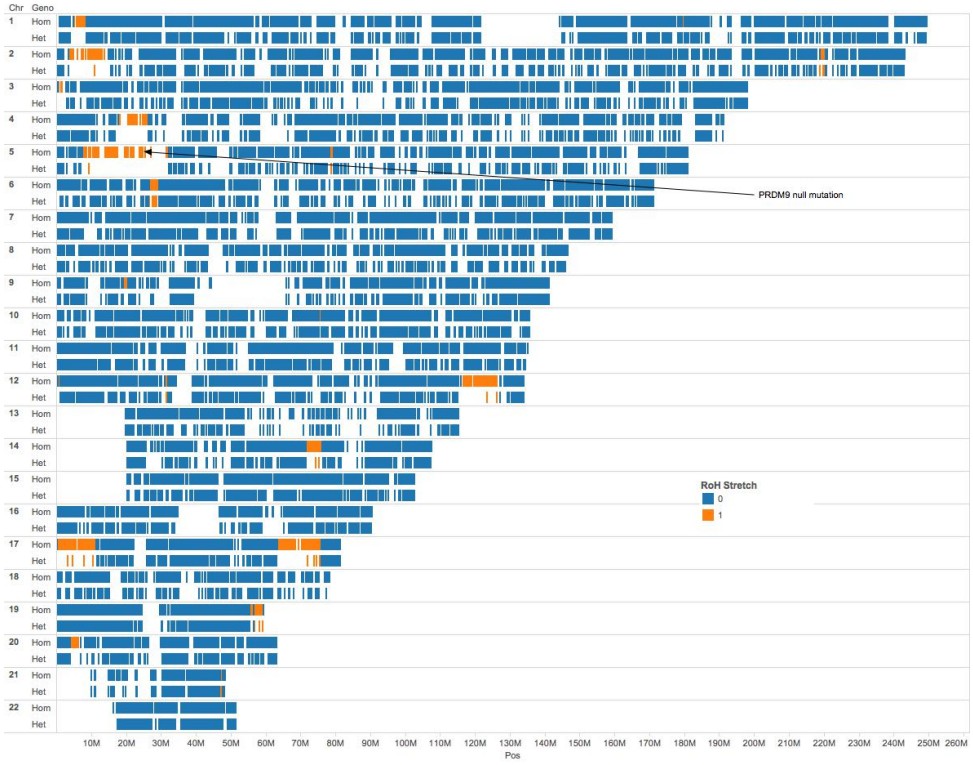


Fig. S7.

Genomic analysis of PRDM9 rhLOF mother and her child. (A) Exome sequencing reads showing PRDM9 chr5:23524525 C / T variant in a control (upper image, genotype CC) sample (upper), and in the mother (lower image, genotype TT). Images from the Integrated Genome Viewer. (B) Sanger dideoxy sequence traces of child (upper trace). This confirms the expected heterozygous genotype and germline transmission of this allele. Middle trace mother, bottom trace control sample. (C) Genomic map of autozygous (identical-by-descent) regions in the mother, showing position of the PRDM9 rhLOF.

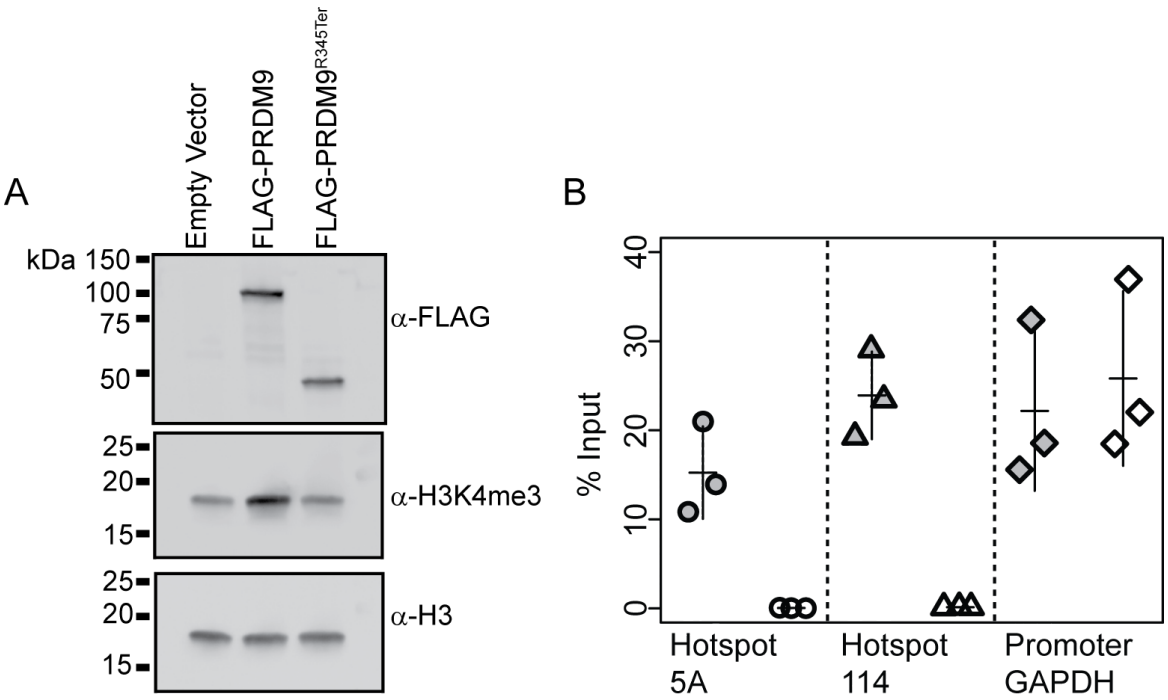


Fig. S8.

Expression of PRDM9 in HEK293 cells. FLAG-tagged *PRDM9* constructs were expressed for 48h in HEK293 cells. (A) Upper panel: the expected protein sizes (full length 110.1kDa, R345Ter 43.2kDa) were confirmed on western blot. Middle panel: expression of mutant *PRDM9* (R345Ter) does not increase global H3K4me3 indicating loss-of-function. Lower panel: histone H3 levels are similar across all samples. (B) R345Ter specifically disrupts *PRDM9*-dependent methylation at hotspots. ChIP for H3K4me3 after expression of *PRDM9* in HEK293 cells followed by qPCR at two recombination hotspots (5A and 114) and PRDM9-independent promoter (*GAPDH*). Filled symbols: full length PRDM9, open symbols R345Ter. Lines indicate mean±SD.

GENE	SEVERE IMPACT	CHR	POS (GRCh37)	REF ALLELE	ALT ALLELE	VALIDATION (of rhLOF genotype)
ANKRD33	splice_acceptor	12	52282350	A	G	yes
BHMT2	splice_acceptor	5	78379449	A	G	yes
BTN3A3	stop_gained	6	26446011	G	A	(unable to design primers / obtain clear PCR band)
C1orf127	stop_gained	1	11008613	G	A	yes
C1QTNF8	stop_gained	16	1143912	G	T	yes
CALHM2	frameshift	10	105209371	CA	C	yes
CD33	frameshift	19	51738926	TC	T	yes
CDK15	stop_gained	2	202744869	C	T	yes
CTB-133G6.1	splice_donor	19	7437958	G	T	yes
CWH43	frameshift	4	49034669	CA	C	yes
DPYD	splice_donor	1	97915614	C	T	yes
EMR2	frameshift	19	14863172	AAG	A	yes
FAM92B	stop_gained	16	85133762	G	A	yes
FLG	frameshift	1	152285076	CACTG	C	(unable to design primers / obtain clear PCR band)
GCA	stop_gained	2	163215655	C	T	yes
GJB2	stop_gained	13	20763490	C	T	yes
HTR2B	stop_gained	2	231973941	T	A	yes
HTR2B	stop_gained	2	231973941	T	A	yes
LIFR	frameshift	5	38530726	CA	C	yes
LINC00935	splice_donor	12	49121343	G	A	yes
LSP1	stop_gained	11	1874388	C	A	yes
MICALCL	frameshift	11	12316171	C	CA	yes
MSRA	stop_gained	8	9912126	C	T	yes
MYO1A	stop_gained	12	57437748	G	A	yes
MYO1A	splice_acceptor	12	57424959	C	A	yes
PLD2	splice_acceptor	17	4711568	G	A	yes
PRDM9	stop_gained	5	23524525	C	T	yes
RP11-293M10.1	frameshift	14	75705824	CA	C	yes

RP11-552I14.1	splice_donor	12	103558217	T	C	yes
SAMD9	stop_gained	7	92732604	G	T	yes
SLC26A10	frameshift	12	58016602	AGCTGCCCAGGATTCTG	A	yes
SLC27A3	splice_donor	1	153750362	G	C	yes
TRDN	stop_gained	6	123539855	A	T	yes
USP45	stop_gained	6	99936089	G	C	yes
ZBBX	splice_donor	3	167083673	C	G	yes
ZBTB24	frameshift	6	109787512	TAG	T	yes
ZSCAN16	stop_gained	6	28097384	C	T	yes

Table S1.
Sanger sequencing validation of variants identified by exome sequencing.

Gene	Number of Individuals with rhLOF	Chr	Pos	REF allele	ALT allele	Comments
C12orf57	2	12	7053817	TGGGT CAGAC GCGGG AAGGC	T	Frameshift not affecting splice site - error in Variant Effect Predictor 79.
CLN3	2	16	28499964	G	A	LOF is relative to a misassigned reading frame for this transcript in GENCODE.
DPM2	1	9	130698730	G	A	Alternate exon in GENCODE, not in RefSeq or UCSC annotation, possible misannotation.
DTNBP1	1	6	15524715	G	A	LOF annotation for only 1 of 4 transcripts containing this variant, extends penultimate exon another 30 amino acids before stop codon, poorly conserved. Possible misannotation.
GHRHR	1	7	31009418	AC	A	Variant not in canonical transcript (which has 3 upstream exons), variant is in non-conserved region of alternative transcripts, possible misannotation.
LRTOMT	2	11	71807767	A	C	Only 2 of 12 GENCODE transcripts include this region as an exon, possible misannotation.
POMGNT1	2	1	46654912	A	T	Canonical transcript is dubious - other transcripts are likely the proper frame, where this is a UTR splice variant.
PUS1	1	12	132416630	A	T	LOF annotation for only 1 of 5 transcripts, other transcripts do not use the first exon containing the variant. Possible misannotation.
NF1	1	17	29706042	G	A	Splice in UTR in all but 1 transcript, likely erroneous transcript.
WWOX	3	16	79245892	GGGGC T	G	LOF annotation for only 1 of 5 long transcripts, likely misassigned reading frame for this transcript in GENCODE.

Table S2.

Variants in OMIM-disease genes erroneously described as LOF due to genome/transcript misannotation.

Gene	Chr	Pos	REF allele	ALT allele	OMIM entry	Relevant Read codes (primary healthcare record)	Is healthcare record compatible with OMIM phenotype?	Additional Comments on OMIM genotype-phenotype association	Additional Comments on genome annotation
AHI1	6	135726088	CT	C	Joubert syndrome-3, 608629	-	No	One report of C-terminal SH3 domain LOF variants being tolerated	-
AKR1D1	7	137792166	AT	A	Bile acid synthesis defect, congenital, 2, 235555	-	No	-	-
ALDH6A1	14	74533476	G	A	Methylmalonate semialdehyde dehydrogenase deficiency, 614105	-	No	Whilst recessive, only missense mutations described. Reduced enzyme activity reported in one subject consistent with LOF.	-
ARMC4	10	28270472	TTC	T	Ciliary dyskinesia, primary, 23, 615451	asthma	Partial	-	-
ATR	3	142178067	A	G	Seckel syndrome 1, 210600	-	No	Only two patients reported for autosomal recessive ATR related disease.	-
ATR	3	142178067	A	G	Seckel syndrome 1, 210600	-	No	Only two patients reported for autosomal recessive ATR related disease.	-
COL1A2	7	94043055	GGTGA	G	Ehlers-Danlos syndrome, cardiac valvular form, 225320	-	No	Relatively few patients reported.	Possibility splice site may be unaffected by variant.
COL6A2	21	47549195	CCT	C	Bethlem myopathy, 158810 Ullrich congenital muscular	-	No	-	Variant only affects some of multiple

					dystrophy 1 254090				transcripts, and is in short homopolymer.
DCHS1	11	6662402	G	GTGT GGTT	Van Maldergem syndrome 1, 601390	-	No	Single published report although multiple affected individuals with different variants.	-
DNAI2	17	72295871	C	T	Ciliary dyskinesia, primary, 9, with or without situs inversus, 612444	allergic rhinitis, asthma, chronic sinusitis, chest infection [multiple entries], primary ciliary dyskinesia, situs inversus	Yes	-	-
DYSF	2	71788881	G	T	Miyoshi muscular dystrophy 1, 254130	genetic syndrome, muscular dystrophy, ECG: Q-T interval abnormal	Yes	-	-
EXPH5	11	108381750	AT	A	Epidermolysis bullosa, nonspecific, autosomal recessive, 615028	-	No	-	-
FBXO7	22	32875119	G	C	Parkinson disease 15, autosomal recessive, 260300	-	No	-	Variant is in a single alternatively spliced transcript. Other variants are coding at this position, or start after this position.
FLG	1	152285076	CACTG	C	Ichthyosis vulgaris, 146700	dermatitis, eczema	Yes	-	-
FLG	1	152276737	TC	T	Ichthyosis vulgaris, 146700	asthma, dermatitis, eczema	Yes	-	-
GJB2	13	20763490	C	T	Deafness, autosomal recessive 1A, 220290	telephone consultation [multiple entries. suggests adequate hearing ability]	No	This variant known to have variable expressivity	-
GJB4	1	35227241	G	A	Erythrokeratodermia variabilis with erythema	dermatitis NOS, rosacea	Partial	Most reports are of heterozygous	-

					gyratum repens, 133200			variants (dominant), only one homozygous (recessive) patient described.	
GYS2	12	21713391	G	GA	Glycogen storage disease 0, liver, 240600	[normal liver function tests]	No	Mild/subclinical phenotypes reported.	-
LIFR	5	38530726	CA	C	Stuve-Wiedemann syndrome/Schwartz-Jampel type 2 syndrome, 601559	-	No	-	Some transcripts have start sites downstream of variant in first exon.
MYO3A	10	26315400	C	T	Deafness, autosomal recessive 30, 607101	hearing test normal	No	Variability reported in age of onset of hearing loss.	-
NEB	2	152544805	C	T	Nemaline myopathy 2, autosomal recessive, 256030	-	No	-	-
NME8	7	37907302	A	G	Ciliary dyskinesia, primary, 6, 610852	hayfever, asthma, rhinitis	Partial	Only single patient case report.	Exon not used in all transcripts.
NPC2	14	74947404	C	T	Niemann-pick disease, type C2, 607625	X-linked centronuclear myopathy [Note different type of myopathy in healthcare record to OMIM entry.]	Yes	-	Nearby in-frame splice site available, which may rescue splicing.
OPA3	19	46032442	G	A	3-methylglutaconic aciduria, type III, 258501	-	No	Suggestion in literature that less deleterious alleles could present as later optic atrophy. This variant truncates ~41aa of 180aa gene.	-
PDE6A	5	149240535	C	CT	Retinitis pigmentosa 43, 613810	-	No	-	Splice acceptor site in last exon.

PLA2G7	6	46679232	C	T	Platelet-activating factor acetylhydrolase deficiency, 614278	-	No	Variable penetrance (of asthma, thrombosis) reported in literature. Despite OMIM annotation, is more of a disease risk modifier than strong Mendelian phenotype.	-
PLA2G7	6	46679232	C	T	Platelet-activating factor acetylhydrolase deficiency, 614278	-	No	Variable penetrance (of asthma, thrombosis) reported in literature. Despite OMIM annotation, is more of a disease risk modifier than strong Mendelian phenotype.	-
PNPLA1	6	36274148	T	A	Ichthyosis, congenital, autosomal recessive 10, 615024	-	No	-	-
RDH5	12	56115279	G	A	Fundus albipunctatus, 136880	-	No	-	-
SAG	2	234243675	C	T	Oguchi disease-1, 258100	-	No	-	-
SAMD9	7	92732604	G	T	Tumoral calcinosis, familial, normophosphatemic, 610455	-	No	-	-
SLC24A1	15	65931948	AC	A	Night blindness, congenital stationary (complete), 1D, autosomal recessive, 613830	no visual symptom	No	Published report from single large pedigree, although there is supportive mouse model.	-
SP110	2	231036485	T	A	Hepatic venoocclusive	-	No	-	Variant is in a

					disease with immunodeficiency, 235550				conserved exon, although exon only used in one of multiple alternatively spliced transcripts.
TMC8	17	76127762	C	CT	Epidermodysplasia verruciformis, 226400	-	No	-	Possible alternative start site downstream of variant, which is in first exon.
TRNT1	3	3188249	T	C	Sideroblastic anemia with B-cell immunodeficiency, periodic fevers, and developmental delay, 616084	[normal full blood count]	No	-	Single alternatively spliced transcript. Non-conserved splice site in middle of exon in other transcripts, possible annotation error.
TRPM1	15	31318399	TC	T	Night blindness, congenital stationary (complete), 1C, autosomal recessive, 613216	electromyogram (EMG) abnormal, impaired vision	Yes	-	-
XDH	2	31625970	G	GC	Xanthinuria, type I, 278300	-	No	-	-
ZBTB24	6	109787512	TAG	T	Immunodeficiency-centromeric instability-facial anomalies syndrome-2, 614069	-	No	-	-

Table S3.

Compatibility of primary healthcare records of Born In Bradford subjects with OMIM recessive genetic diseases for which they have rhLOF genotypes. OMIM autosomal recessive phenotypes only. Each row represents a sequenced individual (i.e. separate individuals if a gene name is present twice). Entries marked [ ] are our interpretation/summary of healthcare record codes. Unoconfirmed or possibly spurious OMIM genotype-phenotype associations, and clearly false genome annotations were previously filtered at the OMIM morbidmap analysis stage or in Table S2, respectively. The two additional comments columns provide further information relevant to the likelihood of true LOF genomic annotation and genotype-phenotype association

	STRING Reference (n=15561)	Born in Bradford LOF all (n=476)	Decode LOF all (n=3769)	Decode LOF "benign" (n=3530)	Decode LOF "pathogenic" (n=416)	Orphanet GOF all (n=106)	Orphanet LOF all (n=458)
Binding p-value, median, [5th-95th quantiles]	10, [0-139]	**9.3×10^{-11} 5.5, [0-61]**	**6×10^{-07} 9, [0-115]**	**2×10^{-06} 9, [0-106]**	0.0105 7, [0-271]	**2.8×10^{-09} 26, [2-152]**	0.0018 14, [0-95]
Reaction p-value, median, [5th-95th quantiles]	0, [0-77]	0.0047 0, [0-39]	0.0152 0, [0-71]	0.0019 0, [0-55]	0.0134 0, [0-543]	**3×10^{-06} 0, [0-52]**	0.0499 0, [0-73]
Activation p-value, median, [5th-95th quantiles]	0, [0-6]	0.533 0, [0-4]	**3.9×10^{-06} 0, [0-4]**	**1.9×10^{-06} 0, [0-4]**	0.3139 0, [0-4]	**1×10^{-17} 1, [0-44]**	**3.3×10^{-13} 0, [0-11]**
Expression p-value, median, [5th-95th quantiles]	0, [0-8]	0.3723 0, [0-6]	**3.6×10^{-05} 0, [0-5]**	**4×10^{-05} 0, [0-5]**	0.0309 0, [0-4]	**1.7×10^{-17} 1, [0-45]**	**1.8×10^{-18} 0, [0-23]**
Catalysis p-value, median, [5th-95th quantiles]	0, [0-9]	**0.0007 0, [0-1]**	**0.0001 0, [0-4]**	**0.0003 0, [0-4]**	0.019 0, [0-1]	**7.9×10^{-08} 0, [0-27]**	0.0001 0, [0-17]
Post Trans Mod p-value, median, [5th-95th quantiles]	0, [0-6]	0.002 0, [0-3]	0.0051 0, [0-4]	0.0119 0, [0-5]	0.00135 0, [0-3]	**6.4×10^{-27} 1.5, [0-31]**	**8.2×10^{-05} 0, [0-9]**
All Interactions p-value, median, [5th-95th quantiles]	14, [0-211]	**3.4×10^{-09} 7, [0-98]**	**4.8×10^{-08} 11, [0-166]**	**4.7×10^{-08} 11, [0-155]**	0.0438 10, [0-815]	**2×10^{-12} 47.5, [4-305]**	**1.1×10^{-06} 24, [1-180]**

Table S4.

A systematic evaluation of the degree of protein-protein interaction in genes carrying Loss of Function (LOF) variants. The following gene sets were compared against genome-wide interactions in the STRING dataset: "Born in Bradford LOF all" (n=476), representing a non-redundant list of genes with observed LOF variants observed in this study in the Born in Bradford cohort. "Decode LOF all" (n=3769), represents all genes with LOF variants in an Icelandic sample(*3*), these genes are further divided into two subgroups

“Decode LOF “benign”” (n=3530), a subset of genes with LOF variants identified in healthy subjects and “Decode LOF “pathogenic”” (n=416), a subset of genes with LOF variants identified in subjects with 1 or more offspring that died before age 15. The final two gene sets, “Orphanet GOF all” (n=106) and “Orphanet LOF all” (n=458), represent genes with pathogenic Mendelian gain of function and loss of function variants respectively reported in the Orphanet rare disease catalogue (www.orpha.net). Bonferroni corrected (42 comparisons) P values <0.0012 shown in bold (Kruskal-Wallis test).

Evaluated individuals and genomic sections / VQSR true positive percentage									
	Genome wide			Just within autozygous regions			Random 176 individuals genome wide		
Pairwise Genotype	90	99	100	90	99	100	90	99	100
Ref-Ref	228,247,662	281,006,032	292,678,961	8,616,089	10,481,646	10,856,209	224,460,477	276,750,800	288,326,547
Het-Het	5,426,895	5,837,611	5,956,070	653	1,391	1,739	1,868,839	2,020,806	2,089,574
Alt-Alt	4,309,160	4,726,190	4,788,443	266,152	287,695	290,382	2,433,917	2,717,644	2,757,343
Missing-Ref	1,244,242	1,543,985	1,644,323	43,713	53,240	56,253	1,527,797	1,938,219	2,081,309
Ref-Missing	1,248,798	1,542,516	1,642,670	46,800	56,065	59,211	1,393,164	1,783,037	1,918,566
Missing-Missing	878,887	1,247,080	1,407,000	21,494	32,627	37,812	586,842	879,125	1,000,909
Missing-Het	420,925	480,144	493,361	355	670	728	188,788	227,049	238,645
Het-Missing	410,839	468,232	481,478	392	655	730	170,117	206,356	218,024
Missing-Alt	186,400	224,312	231,504	11,400	13,354	13,579	238,773	282,135	290,149
Alt-Missing	181,757	217,983	224,886	11,579	13,487	13,745	223,521	263,617	270,946
Ref-Het	72,030	92,965	99,693	235	492	546	3,016,207	3,293,345	3,347,482
Het-Ref	70,691	92,657	99,580	300	579	636	3,017,736	3,299,219	3,353,016
Het-Alt	29,063	32,816	33,459	1,178	1,316	1,333	1,092,254	1,163,195	1,173,963
Alt-Het	27,673	31,214	31,834	835	948	956	1,092,330	1,162,645	1,173,613
Ref-Alt	3,306	3,670	3,718	4	5	6	726,823	784,369	792,954
Alt-Ref	3,013	3,399	3,466	2	2	2	723,756	779,245	787,406

Table S5.

Pairwise genotypes for duplicate samples across evaluated individuals and regions. Pairwise genotypes (Ref, Het, Alt) across 176 duplicate pairs that were evaluated genome wide and just within the autozygous sections along with those from 176 unrelated (random) pairs. VQSR true positive percentage reflect the portion of true positive sites left (at VQSR= 90, 99, 100).

	BEB	CDX	CEU	CHB	CHS	ESN	FIN	GBR	GIH	GWD	IBS	ITU	JPT	LWK	MSL	PJL	STU	TSI	YRI
BB	**0.780** (0.074)	**0.820** (0.067)	**0.840** (0.077)	**0.810** (0.075)	0.940 (0.069)	**0.820** (0.073)	0.960 (0.077)	0.880 (0.067)	0.890 (0.066)	**0.860** (0.068)	**0.840** (0.065)	**0.790** (0.076)	**0.760** (0.066)	**0.840** (0.070)	**0.830** (0.068)	**0.790** (0.081)	**0.840** (0.078)	**0.780** (0.069)	0.890 (0.076)
BEB		1.080 (0.075)	1.050 (0.074)	1.060 (0.073)	**1.200** (0.070)	1.070 (0.071)	**1.290** (0.068)	**1.160** (0.069)	1.060 (0.071)	1.100 (0.074)	1.070 (0.068)	1.070 (0.077)	0.980 (0.070)	1.080 (0.075)	1.070 (0.073)	1.040 (0.067)	1.040 (0.066)	1.010 (0.073)	1.100 (0.073)
CDX			0.970 (0.070)	0.980 (0.070)	1.120 (0.067)	1.000 (0.073)	**1.200** (0.067)	1.070 (0.069)	0.980 (0.070)	1.030 (0.065)	0.990 (0.074)	0.980 (0.074)	0.910 (0.072)	1.010 (0.072)	1.000 (0.071)	0.960 (0.073)	0.960 (0.066)	0.940 (0.068)	1.030 (0.071)
CEU				1.010 (0.067)	**1.150** (0.064)	1.020 (0.071)	**1.240** (0.070)	1.100 (0.068)	1.010 (0.076)	1.050 (0.072)	1.020 (0.069)	1.010 (0.069)	0.940 (0.068)	1.030 (0.077)	1.020 (0.067)	0.980 (0.069)	0.980 (0.072)	0.960 (0.071)	1.050 (0.067)
CHB					**1.140** (0.069)	1.010 (0.076)	**1.220** (0.068)	1.090 (0.079)	0.990 (0.067)	1.040 (0.070)	1.010 (0.064)	0.990 (0.074)	0.930 (0.072)	1.020 (0.075)	1.010 (0.072)	0.970 (0.076)	0.980 (0.074)	0.960 (0.070)	1.040 (0.072)
CHS						0.900 (0.071)	1.070 (0.072)	0.950 (0.066)	0.870 (0.071)	0.920 (0.070)	0.880 (0.077)	0.870 (0.076)	**0.810** (0.064)	0.910 (0.071)	0.900 (0.072)	**0.850** (0.066)	**0.860** (0.070)	**0.840** (0.067)	0.930 (0.074)
ESN							**1.190** (0.075)	1.070 (0.069)	0.990 (0.075)	1.040 (0.070)	1.000 (0.070)	0.990 (0.066)	0.930 (0.071)	1.010 (0.066)	1.010 (0.068)	0.970 (0.068)	0.970 (0.073)	0.950 (0.071)	1.040 (0.066)
FIN								0.890 (0.072)	**0.810** (0.082)	0.860 (0.077)	**0.820** (0.080)	**0.810** (0.067)	**0.760** (0.070)	0.850 (0.076)	**0.840** (0.069)	**0.800** (0.075)	**0.800** (0.073)	**0.790** (0.072)	0.870 (0.071)
GBR									0.910 (0.080)	0.960 (0.067)	0.920 (0.070)	0.910 (0.063)	0.850 (0.080)	0.950 (0.070)	0.940 (0.069)	0.890 (0.074)	0.890 (0.069)	0.870 (0.069)	0.970 (0.069)
GIH										1.050 (0.077)	1.020 (0.077)	1.000 (0.073)	0.930 (0.073)	1.030 (0.069)	1.020 (0.070)	0.970 (0.070)	0.980 (0.073)	0.960 (0.068)	1.050 (0.071)
GWD											0.970 (0.073)	0.950 (0.065)	0.900 (0.073)	0.980 (0.074)	0.970 (0.065)	0.930 (0.067)	0.930 (0.070)	0.920 (0.071)	1.000 (0.069)
IBS												0.980 (0.071)	0.920 (0.073)	1.020 (0.072)	1.000 (0.075)	0.960 (0.072)	0.960 (0.074)	0.940 (0.074)	1.040 (0.077)
ITU													0.930 (0.075)	1.030 (0.063)	1.020 (0.074)	0.970 (0.075)	0.980 (0.066)	0.960 (0.071)	1.050 (0.073)
JPT														1.090 (0.066)	1.080 (0.075)	1.050 (0.068)	1.050 (0.075)	1.030 (0.066)	1.110 (0.061)
LWK															0.990 (0.078)	0.950 (0.065)	0.950 (0.075)	0.940 (0.074)	1.020 (0.073)
MSL																0.960 (0.065)	0.960 (0.069)	0.940 (0.068)	1.030 (0.066)
PJL																	1.000 (0.074)	0.980 (0.081)	1.070 (0.056)
STU																		0.980 (0.070)	1.070 (0.073)
TSI																			1.090 (0.073)

Table S6.

Pairwise $R_{A,B}$ values across different populations. Numbers in bold reflect the pairwise $R_{A,B}$ values for LOF variants normalized by the synonymous variant counts while the jacknife standard errors are shown in brackets. Highlighted in blue are comparisons that are at least 2 standard errors away from 1 (neutral expectation) and in orange are comparisons that are 3 standard errors away.

	SNPS phased	DNA in molecules >20kb	GEMs detected	N50 phase block	Mean depth
NA12878	96.2%	89.1%	247924	16674432 bp	33.9 X
NA12882	97.8%	85.4%	245880	9346942 bp	28.2 X
PRDM9 knockout mother	92.6%	70.4%	235515	586148 bp	29.9 X
Her son	90.7%	27.9%	223067	157393 bp	26.2 X

Table S7.

10XGenomics summary statistics on sequenced samples.

	Allelic combinations															
Mother 1	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0
Mother 2	0	0	1	1	0	0	1	1	1	1	0	0	1	1	0	0
Child mat	0	0	1	1	0	0	1	1	0	0	1	1	1	1	0	0
Child pat	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Table S8.

Allelic combinations of sites informative about recombination in mother child duos. Numbers show the different alleles (0-reference, 1-alternate) seen in the parent and child for bi-allelic sites. The first 8 configurations show transmission of the chromosome Mother 2 while the next 8 show transmission of chromosome Mother 1. Sites that are informative when only the mother is phased are shown in orange, while additional sites that are informative when the child is also phased so as to identify the maternal chromosome are shown in yellow.